Annotation and Multi-modal Methods for Quality Assessment of Multi-party Discussion

Tsukasa Shiota
Kyushu Institute of Technology
680-4 Kawazu Iizuka Fukuoka JAPAN

Kazutaka Shimada
Kyushu Institute of Technology
680-4 Kawazu Iizuka Fukuoka JAPAN
shimada@ai.kyutech.ac.jp

Abstract

Discussion quality assessment tasks have recently attracted significant attention in natural language processing. However, there have been few studies on challenging such tasks, with a focus on synchronous discussions. In this study, we annotate quality scores to each discussion in an existing multi-modal multi-party discussion corpus. Furthermore, we propose some quality assessment methods with multi-modal inputs. As the results show, attention-based long short-term memory (LSTM) with multi-modal inputs produces the best performance for the "Effectiveness" criterion whereas text information has an important role in the "Reasonableness."

1 Introduction

In recent years, problem-based and cooperative learning have been attracting attention as a means of skills training, such as communication skills, in education. One educational training approach is a group discussion, which involves debate and consensus-building. Introducing this learning approach to a classroom requires a great deal of effort to evaluate and provide feedback on the abilities and achievements of all groups and individuals from various perspectives because several discussion groups usually exist in a single class at the same time. Furthermore, an assessment is a difficult task because there are no correct answers regarding discussions in general. Moreover, quantitative and objective evaluations are difficult. Therefore, an automatic assessment, such as a visualization of the discussion state and a judgment of the discussion score, is a desirable and valuable task for education, examinations through discussion, and so forth. It will be possible to reduce the burden of evaluation activities on the evaluators.

One of the educational applications in natural language processing is automated essay scoring (AES) (Ke and Ng, 2019) as an argument quality evaluation. However, the structure of a spoken discussion, which is our target in this paper, is not as clear as that of a written discussion. In addition, in spoken discussions, both verbal and non-verbal information have important roles in understanding and evaluating the discussion. Mukawa et al. (2018) have reported that non-verbal features, such as gestures and an interval of utterances, have a powerful effect on group discussions.

In this study, we annotate several quality assessment criteria and scores to a multi-modal multi-party discussion corpus. The language used is Japanese, and the corpus is freely available[1]. In addition, we propose the use of machine-learning-based methods, such as a support vector machine (SVM) and neural networks, and then evaluate the methods using multi-modal inputs. In the experiment, we discuss the relationships between the assessment criteria and input modalities.

2 Related work

There are some dialogue and meeting corpora (Carletta, 2007; Janin et al., 2003). Some face-to-face discussion corpora have been also developed. Zhang et al. (2016) have constructed a corpus through a competitive debate format. They reported that their method predicts the winner of each debate at a rate of approximately 60%. Hayashi et al. (2015) have developed a group discussion interaction corpus to evaluate five communication skills. This corpus contains not only transcriptions but also speech, gaze, head motions, and poses using certain devices. Olshefski et al. (2020) have constructed a discussion tracker corpus in an educational environment. The corpus consists of 29 multi-party discussions. Yamamura et al. (2016) have constructed a corpus for a discussion summarization. The corpus consists of 9 discussions by four participants.

[1] http://www.pluto.ai.kyutech.ac.jp/~shimada/resources.html#kyutechDB

As mentioned in Section 1, many studies and corpora of asynchronous and written texts exist, such as essay writing (Ke and Ng, 2019). Some researchers have recently studied interactions between participants during discussions. For example, Okada et al. (2016) have annotated communication skill scores on the MATRICS corpus (Nihei et al., 2014). They also proposed a multi-modal prediction model for such a task. In addition, Avci and Aran (2016) and Murray and Oertel (2018) have proposed performance prediction models by using features extracted from the states of the discussions and the participants. In this paper, we also introduce multi-modal features to our quality assessment method.

3 Dataset

Our purpose in this paper is to assess a quality of a multi-party discussion. For the purpose, we need a discussion corpus. In this paper, we utilize the corpus that was constructed by (Shiota and Shimada, 2020), namely the Kyutech Debate corpus. It is freely available on their website[2]. This section describes their corpus first and then our annotation process for our purpose.

In the Kyutech Debate corpus, two people in a group first debated an issue from both positive and negative standpoints, and the two groups then came to a consensus through compromise. The first (debate) and second (consensus-building) parts were each 20 min in length. The discussions of five groups were recorded, with 200 min of discussions as a whole. The corpus consisted of 7,449 utterances that were transcribed[3], body key-points determined by OpenPose[4], facial landmarks determined by OpenFace[5], and the speech features analyzed using Surfboard[6].

In this paper, we newly add quality assessment scores for the corpus. In general, participants need to discuss various topics in the case of debating/consensus-building of an issue. Hence, we extract topic-based segments (hereinafter referred to as "discussion segments") and regard them as the target units of a quality assessment. We referred to the topic segmentation manual (Xu et al., 2005) of the AMI corpus, which is a popular conversation corpus. As a result, we obtained 178 segments from the Kyutech Debate corpus.

Next, we created criteria based on the theory of computational quality assessment of natural language arguments (Wachsmuth et al., 2017b) and conducted a grading process. According to the classification defined by the above study, the quality of an argument can be evaluated through two main criteria, "Reasonableness (Re)" and "Effectiveness (Ef)," and their sub-criteria. The sub-criteria of Re are "Global Acceptability (GA)," "Global Relevance (GR)," and "Global Sufficiency (GS)." The sub-criteria of Ef are "Credibility (Cr)," "Emotional appeal (Em)," "Clarity (Cl)," "Appropriateness (Ap)," and "Arrangement (Ar)." Table 1 provides a description of each criterion based on the previous study.

Three workers, who were graduate students and not related to this work in our laboratory, were assigned to each discussion segment. Given the transcription and video data of a discussion segment, they judged the quality of each segment on the basis of the more detailed explanation provided in Table 1. The first step was to rate each sub-criterion as low (L), middle (M), or high (H), and then determine the score of the main criteria on the basis of the score distribution of the sub-criteria: very low (VL), L, M, H, or very high (VH).

The reliability of the annotated main and sub-criteria was confirmed by calculating the agreement rate. Table 2 shows Krippendorff's α coefficient for each criterion. This coefficient is a continuous value of between -1 and 1, which can be used to calculate the rate of agreement for all scales. The values in the table are not always high. The result denotes that the annotation task is inherently difficult. As a similar study, Wachsmuth et al. (2017a) also reported an annotation process and the result using the same scheme for the written text. In their study, the Klippendorff α of crowd workers ranged from -0.27 to 0.53. In other words, the result was also low. Moreover, the values of some criteria by Wachsmuth et al. (2017a) dropped below zero. On the other hand, such results did not appear on our annotation. Therefore, our annotated data contain a better point than the previous study. In other words, our data are a better-than-random chance although the previous work contained a result that was less than a ran-

[2] http://www.pluto.ai.kyutech.ac.jp/~shimada/resources.html

[3] Transcription units were based on a 0.2 seconds interval.

[4] https://github.com/CMU-Perceptual-Computing-Lab/openpose

[5] https://github.com/TadasBaltrusaitis/OpenFace

[6] https://github.com/novoic/surfboard

Criterion	Explanation
Re	Does the argumentation satisfy GA, GR, and GS?
GA	Does the target audience accept both the consideration of the stated arguments regarding the issue and the way in which they are stated?
GR	Does it contribute to the resolution of the issue?
GS	Does it adequately rebut the counterarguments by properly anticipating them?
Ef	Does the argumentation satisfy Cr to Ar?
Cr	Does it convey arguments and is it similar in such a way that it makes the author worth considering?
Em	Were emotions elicited to make the target audience more open to the author's arguments?
Cl	Was the argument correct and widely unambiguous?
Ap	Did the language used support the credibility?
Ar	Were the issue, arguments, and their conclusion presented in the correct order?

Table 1: Definitions of the Quality Dimensions, based on Wachsmuth's study.

Re	GA	GR	GS	-	-
0.151	0.087	0.029	0.128	-	-
Ef	Cr	Em	Cl	Ap	Ar
0.135	0.032	0.038	0.017	0.076	0.155

Table 2: Krippendorff's α of each criterion.

dom chance. Although this annotation is a complicated task, it is necessary to improve the agreement as one future work. As one of our contributions, we will open the annotated data on the web.

4 Quality Assessment Method

In this section, we describe our quality assessment method for the dataset introduced in Section 3. First, we define the quality assessment task and then propose four models based on the SVM and neural networks.

4.1 Task Definition

A discussion segment S consists of a sequence of utterance vectors $U = \{\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_N\}$. Here, N is the number of utterances in a segment, and $\boldsymbol{u}_i$ is a vector of the i-th utterance in S. The task in this paper is to predict the class labels of each criterion from a sequence U with verbal and non-verbal information. Here, the class labels are L, M, and H, as described in Section 3. Owing to the limited number of instances that belong to each class, VL is merged with L, and VH is merged with H. In other words, the task is a classification task with three class labels for the criteria in Section 3.

Here, $\boldsymbol{u}_i$ is expressed as follows:

$$\boldsymbol{u}_i = [\boldsymbol{sp}_i; \boldsymbol{t}_i; \boldsymbol{b}_i; \boldsymbol{f}_i; \boldsymbol{a}_i] \quad (1)$$

where $[\cdot;\cdot]$ denotes the concatenation of the vectors.

In addition, $\boldsymbol{sp}_i$ denotes whether the speaker of the i-th utterance is different from the speaker of the $i-1$-th utterance. In other words, it is a binary feature, i.e, the speaker is the same (0) or different (1).

Moreover, $\boldsymbol{t}_i$ is a vector from text information of the i-th utterance. For the $\boldsymbol{t}_i$, we use BERT (Devlin et al., 2019). We apply the CLS token (768 dimensions) on the 11th layer from a Japanese BERT developed by Tohoku University[7].

Here, $\boldsymbol{b}_i$ is a vector from the body information of the i-th utterance. It consists of the average and standard deviation of (x, y) values of the nose, neck, right shoulder, right elbow, right wrist, right eye, right ear, left shoulder, left elbow, left wrist, left eye, and left ear from OpenPose (a total of 48 dimensions).

In addition, $\boldsymbol{f}_i$ is a vector from facial information of the i-th utterance. It consists of the average and standard deviation of the facial and eye points (x, y), gaze direction, head location, and head direction. In addition, it contains the presence of facial action units (AUs). OpenFace extracts these values, and the number of dimensions is 586.

Finally, $\boldsymbol{a}_i$ is a vector from audio information of the i-th utterance. It consists of the minimum, maximum, average, and standard deviation of 13 MFCC, the RMS, the fundamental frequency, and the spectral centroid. In addition, it contains the Jitter and Shimmer values. Surfboard extracts these values, and the number of dimensions is 72.

[7] https://github.com/cl-tohoku/bert-japanese

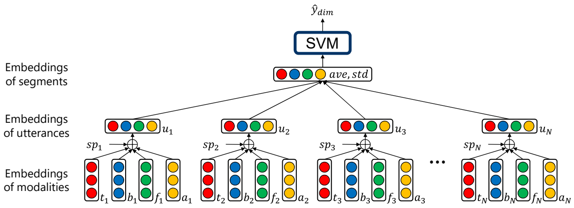

Figure 1: Method based on SVM.

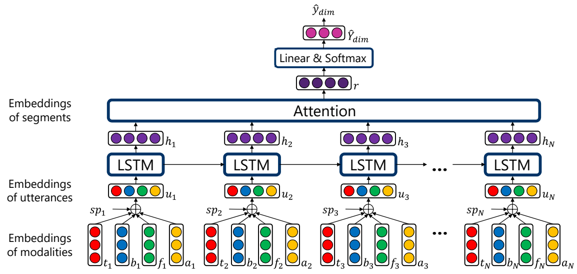

Figure 2: Attention-based LSTM.

4.2 SVM

As one of the simplest models, we apply an SVM (Vapnik, 2013) to the task. Because an SVM cannot handle sequence information directly, we compute the average and standard deviation of each vector in the time sequence and use the values as the vector of each discussion segment. We estimate each quality assessment label $\hat{y}_{dim}$ by using the model with the vector. Figure 1 shows an overview of this method.

4.3 LSTM

As mentioned above, the SVM-based method cannot handle the utterance sequence information well. Therefore, as a suitable model for sequence information, we use LSTM for this task.

Given an input $\boldsymbol{u}_i$, the units of LSTM are computed as follows:

$$\boldsymbol{h}_i = LSTM(\boldsymbol{u}_i, \boldsymbol{h}_{i-1}, \boldsymbol{c}_{i-1}) \tag{2}$$

After the computation for all utterances, LSTM obtains the final state of a discussion segment $\boldsymbol{h}_N$. In this paper, we regard $\boldsymbol{h}_N$ as the embeddings of the discussion segment. We calculate a probability distribution $\hat{Y}_{dim}$ using the softmax function.

$$\hat{Y}_{dim} = \mathrm{softmax}(\boldsymbol{W}_s \boldsymbol{h}_N + \boldsymbol{b}_s), \tag{3}$$

where $\boldsymbol{W}_s$ and $\boldsymbol{b}_s$ are parameters in the learning process, respectively. In addition, $\mathrm{softmax}()$ is the softmax function. Finally, we select the label with the maximum probability ($\hat{y}_{dim}$).

$$\hat{y}_{dim} = \arg\max_{y_{dim}} \hat{Y}_{dim} \tag{4}$$

4.4 Attention-based LSTM

By using the LSTM, we can capture the sequence information of the utterances. However, discussion segments often contain non-important utterances for a quality assessment task, e.g., a nod. Therefore, we introduce attention mechanisms to the LSTM-based method, such as the models from (Wang et al., 2016) and (Zhou et al., 2016).

First, in the same way as the LSTM-based approach, we compute $\boldsymbol{h_i}$ of i. We then compute the weight a_i of $\boldsymbol{h_i}$ by

$$m_i = \boldsymbol{\omega}^T \tanh(\boldsymbol{h}_i), \tag{5}$$

$$a_i = \frac{\exp(m_i)}{\sum\limits_{j=1}^{N} \exp(m_j)}, \tag{6}$$

where $\boldsymbol{\omega}^T$ is a parameter, and $\exp()$ is the exponent function. Next, we obtain the final state $\boldsymbol{h}^*$ by using the summation of hidden layers $\boldsymbol{h}_i$ weighted by a_i.

$$\boldsymbol{r} = \sum_{i=1}^{N} a_i \boldsymbol{h}_i, \tag{7}$$

$$\boldsymbol{h}^* = \tanh(\boldsymbol{r}), \tag{8}$$

where $\tanh()$ is the hyperbolic tangent function. We regard $\boldsymbol{h}^*$ as the embeddings of the segment and calculate $\hat{Y}_{dim}$.

$$\hat{Y}_{dim} = \mathrm{softmax}(\boldsymbol{W}_s \boldsymbol{h}^* + \boldsymbol{b}_s) \tag{9}$$

Finally, we select the label with the maximum probability ($\hat{y}_{dim}$). Figure 2 shows an overview of this method.

4.5 Hierarchical LSTM

By using an LSTM and attention mechanisms, we can handle the state of an utterance sequence. However, $\boldsymbol{t}_i$ does not directly handle the word sequence in an utterance. We therefore incorporate word sequence information with the LTSM-based model similarly to the approach by (Tran et al., 2017).

We compute $\boldsymbol{h}_{i,j}^{Uttr}$ with $\boldsymbol{w}_i$ as follows:

$$\boldsymbol{h}_{i,j}^{Uttr} = LSTM^{Uttr}(\boldsymbol{w}_{i,j}, \boldsymbol{h}_{i,j-1}^{Uttr}, \boldsymbol{c}_{i,j-1}^{Uttr}), \tag{10}$$

$$\boldsymbol{w}_i = \boldsymbol{h}_{i,M_i}^{Uttr}, \tag{11}$$

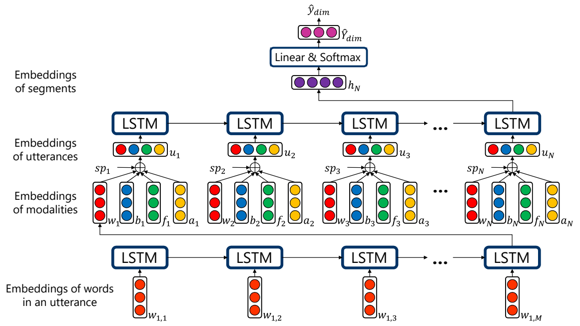

Figure 3: Hierarchical LSTM.

where $LSTM^{Uttr}()$ is an encoder of the sequence of word vectors, and $\boldsymbol{w}_{i,j}$ is the vector of the j-th word in the i-th utterance in a segment S. The vector is also extracted from the 11th layer of BERT. Here, $\boldsymbol{h}_{i,j}^{Uttr}$ is the hidden layer of the LSTM at (i, j), $\boldsymbol{c}_{i,j}^{Uttr}$ is the memory cell of the LSTM at (i, j), and M_i is the number of words in the i-th utterance. Hence, the vector $\boldsymbol{u}_i$ of this method is as follows:

$$\boldsymbol{u}_i = [\boldsymbol{sp}_i; \boldsymbol{w}_i; \boldsymbol{b}_i; \boldsymbol{f}_i; \boldsymbol{a}_i] \tag{12}$$

After that, similarly to the LSTM-based method described in Section 4.3, we also obtain the final state $\boldsymbol{h}_i^{Hier}$, $\hat{Y}_{dim}$, and the class label with the maximum probability.

$$\boldsymbol{h}_i^{Hier} = LSTM^{Hier}(\boldsymbol{u}_i, \boldsymbol{h}_{i-1}^{Hier}, \boldsymbol{c}_{i-1}^{Hier}) \tag{13}$$

$$\hat{Y}_{dim} = \text{softmax}(\boldsymbol{W}_s \boldsymbol{h}_N^{Hier} + \boldsymbol{b}_s) \tag{14}$$

Figure 3 shows an overview of this method.

5 Experiment

5.1 Setting

As mentioned in Section 4.1, we merged the VH class with H and the VL class with L. Hence, the task in this paper is a three-class classification, namely, L, M, and H. The targets of the classification are the two criteria with relatively high Krippendorff α values listed in Table 2: Re (Reasonableness) and Ef (Effectiveness). The statistics are shown in Table 3.

We applied the L2 norm cross-entropy loss as the loss function for the neural network-based methods. We used the SGD (Bottou, 1991) with Momentum (Qian, 1999) ($\alpha = 0.95$) as the optimizer. For the hyperparameters, the size of hidden layers was 500 dimensions, the batch size was 32, the number of epochs was 50, the learning rate was 0.01, the drop-out rate was 0.2, and the decay factor was 0.001.

Criterion	L	M	H
Re	13	89	76
Ef	9	97	72

Table 3: Distribution of each class of two target criteria.

Our dataset is small, namely, 178 segments from 10 discussions. We divided the dataset into eight discussions for the training, one discussion for the development, and one discussion for the test. We then evaluated each method based on a 10-fold cross-validation of the discussion level. We calculate the average F-scores for each criterion, i.e., Re and Ef, based on the cross-validation. For the robustness of the results, we conducted this evaluation five times and then calculated the average values of the five evaluations.

5.2 Results and Discussion

Table 4 shows the experiment results. Here, T, B, F, and A denote the text, body information, facial information, and audio information modalities. The combination of each letter denotes the combination of modalities. For example, TB denotes the combination of text and body information as the input of each method. Hence, TBFA, on the left-most side of the table, denotes the method with all modalities. Boldface denotes the best score among the modality combinations. The underlined values denote the best values of uni-modal, bi-modal, and multi-modal inputs. For example, 0.398, 0.459, and 0.399 are the best scores of TB, TF, and TA, namely, bi-modal inputs. The best score of the bi-modal setting is 0.459 by the TF input. The scores with $*$ denote the best scores for each criterion.

For the Re criterion, multi-modal inputs were not always effective for the classification. However, there were no significant changes in the results when the input modalities were expanded. The best score was 0.459, achieved by hierarchical LSTM (H-LSTM) with text and facial information. However, the difference between H-LSTM and SVM with text only was slight (0.008). Moreover, H-LSTM is a method that can handle word information directly, as compared with LSTM and attention-based LSTM (A-LSTM). Here, recall that the Re criterion consists of the acceptability,

Criteria	Model	F1-Score							
		T	TB	TF	TA	TBF	TBA	TFA	TBFA
Re	SVM	**0.451**	0.338	0.337	0.343	0.333	0.317	0.320	0.340
	LSTM	0.387	**0.398**	0.392	0.380	0.410	0.379	0.388	0.360
	A-LSTM	0.412	0.392	0.387	**0.399**	0.371	0.359	**0.398**	0.398
	H-LSTM	0.359	0.354	**0.459***	0.388	**0.415**	**0.391**	0.370	**0.405**
Ef	SVM	**0.459**	0.382	0.383	0.436	0.384	0.392	0.406	0.379
	LSTM	0.428	**0.478**	**0.438**	**0.476**	**0.467**	**0.472**	**0.486**	0.435
	A-LSTM	0.433	0.470	0.426	0.468	0.450	0.396	0.444	**0.490***
	H-LSTM	**0.459**	0.433	0.416	0.379	0.414	0.440	0.431	0.451

Table 4: Experiment results of four methods with a combination of four modalities.

relevance, and sufficiency of the discussions. In other words, it is related to the content of each discussion. Therefore, non-verbal information is less likely to contribute to improving the accuracy. From these results, we concluded that text information is the most important factor for the Re criterion.

For the Ef criterion, the combination of modalities improved the F-scores except for the SVM-based method. The best F-score was produced by A-LSTM with all modalities (0.490). The Ef criterion is based on the receivers' emotions during the discussions, such as credibility and emotional appeal. In addition, it contains clarity and appropriateness in the discussion. In general, we utilize eye contact (addressing) and body language to clearly convey a message and elicit sympathy. In other words, not only text but also actions, expressions, and the tone of voice of the speakers have an important role for the Ef criterion. From these results, we conclude that incorporating both verbal and non-verbal modalities leads to an improvement of the estimation of this criterion.

One simple method for predicting the label of a criterion is to use the majority label. In this dataset, this is label M for both criteria (Re and Ef) from Table 3. However, note that the distribution of labels in each discussion is uniform. In other words, there is a situation in which most of the labels in a discussion are H, although another discussion contains as many instances of label M as label H. In fact, the F1-scores of the majority selection based on the same calculation approach described in Section 5.1 were 0.333 for Re and 0.384 for Ef[8]. These values were lower than most of the F-scores in Table 4. This result shows the effectiveness of our methods.

[8]Note that these values cannot be calculated from Table 3 because of a lack of label distribution for each discussion.

6 Conclusions

In this paper, we annotated quality assessment scores for an automatic discussion evaluation to a multi-party conversation corpus. We proposed four machine learning methods for the task: SVM-based, LSTM-based, attention-based LSTM, and hierarchical LSTM methods. We used not only text but also non-verbal information, namely, multi-modal inputs.

We evaluated the methods using a 10-fold cross-validation for two criteria at the discussion level, namely, Re and Ef, in the corpus. For Re, the hierarchical LSTM with text and facial information obtained the best F-score. In addition, the SVM with only text information obtained a good result. For this criterion, text information has the most important role because it is related to the content of the discussions. For Ef, the attention-based LSTM with all modalities produced the best F-score. For this criterion, various inputs are essentially suitable because it is related to the impression and emotion of the speakers and receivers in the discussion. However, the F1-scores are insufficient (0.459 for Re and 0.490 for Ef). Improving the method using other information, such as knowledge graphs (Al-Khatib et al., 2020), is an important area of future work.

We annotated several quality assessment criteria to an existing discussion corpus. However, the size of the corpus is not large. Annotation to other corpora is an important task. An improvement of the agreement of each criterion will be also an important future research area, although it is essentially a difficult task.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 20K12110.

References

Khalid Al-Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-End Argumentation Knowledge Graph Construction. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 7367–7374.

Umut Avci and Oya Aran. 2016. Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction. *IEEE Transactions on Multimedia*, 18(4):643–658.

Léon Bottou. 1991. Stochastic Gradient Learning in Neural Networks. In *Proceedings of Neuro-Nîmes 91*.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Jacob Devlin, Ming-Wei Chang, and and Kristina Toutanova Kenton Lee. 2019. BERT: Pre-training of Deep Bidirectional Ttransformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, page 4171–4186.

Yuki Hayashi, Fumio Nihei, Yukiko I. Nakano, Hung-Hsuan Huang, and Shogo Okada. 2015. Development of Group Discussion Interaction Corpus and Analysis of the Relationship with Personality Traits. *IPSJ Journal*, 56(4):1217–1227.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The ICSI Meeting Corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 364–367.

Zixuan Ke and Vincent Ng. 2019. Automated Essay Scoring: A Survey of the State of the Art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6300–6308.

Naoki Mukawa, Tomohiro Nakayama, Hiroko Tokunaga, Junji Yamato, and Naomi Yamashita. 2018. Analysis of Verbal / Nonverbal Expressions of Speakers and Evaluators' Evaluations in Group Discussions. *The IEICE Transactions on Information and Systems*, J101-D(2):284–293.

Gabriel Murray and Catharine Oertel. 2018. Predicting Group Performance in Task-Based Interaction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, pages 14–20.

Fumio Nihei, Yukiko Nakano, Yuki Hayashi, Hung-Hsuan Hung, and Shogo Okada. 2014. Predicting Influential Statements in Group Discussions using Speech and Head Motion Information. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI)*, pages 136–143.

Shogo Okada, Yoshihiko Ohtake, Yukiko Nakano, Hayashi Yuki, Hung-Hsung Huang, Yutaka Takase, and Katsumi Nitta. 2016. Estimating Communication Skills Using Dialogue Acts and Nonverbal Features in Multiple Discussion Datasets. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*, pages 169–176.

Christopher Olshefski, Luca Lugini, Ravneet Singh, Diane Litman, and Amanda Godley. 2020. The Discussion Tracker Corpus of Collaborative Argumentation. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 1033–1043.

Ning Qian. 1999. On the Momentum Term in Gradient Descent Learning Algorithms. *Neural Networks*, 12(1):145 – 151.

T. Shiota and K. Shimada. 2020. The Discussion Corpus toward Argumentation Quality Assessment in Multi-Party Conversation. In *Proceedings of LTLE*, pages 280–283.

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017. A Hierarchical Neural Model for Learning Sequences of Dialogue Acts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 428–437.

Vladimir Vapnik. 2013. *The Nature of Statistical Learning Theory*. Springer science & business media.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation Quality Assessment: Theory vs. Practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 250–255.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 176–187.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 606–615.

Weiqun Xu, Jean Carletta, Jonathan Kilgour, and Vasilis Karaiskos. 2005. Coding Instructions for Topic Segmentation of the AMI Meeting Corpus Version 1.1.

Takashi Yamamura, Kazutaka Shimada, and Shintaro Kawahara. 2016. The Kyutech Corpus and Topic Segmentation Using a Combined Method. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR)*, pages 95–104.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Flow in Oxford-style Debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 136–141.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 207–212.