# Analyzing Collaborative Learning Process by Deep Learning Methods: A Multi-Dimensional Coding Scheme with an Assessment Model

Taketoshi Inaba, Chihiro Shibata
Graduate School of Bionics, Computer and Media Sciences
Tokyo University of Technology
Tokyo, Japan
email: {inaba, shibatachh}@stf.teu.ac.jp

Kimihiko Ando
Cloud Service Center
Tokyo University of Technology
Tokyo, Japan
email:ando@stf.teu.ac.jp

***Abstract*—In computer-supported collaborative learning research, it may be a significantly important task to figure out guidelines for carrying out an appropriate scaffolding by extracting indicators for distinguishing groups with poor progress in collaborative process upon analyzing the mechanism of interactive activation. And for this collaborative process analysis, labelling for appropriately representing properties of each contribution (coding) and statistical analysis are often adopted as a method. But as far as this paper is concerned, it tries to automate this huge laborious coding work with deep learning technology. In its previous research, supervised data was prepared for deep learning based on a coding scheme consisting of 16 labels according to speech acts. In this paper, with a multi-dimensional coding scheme with five dimensions newly designed aiming at analyzing collaborative learning process more comprehensively and multilaterally, an automatic coding is performed by deep learning methods and its accuracy is verified. The results indicate with certainty that we can introduce this model to authentic educational settings and that even for large classes with many students, we can perform real-time monitoring of learning process or ex-post analysis of big educational data. However, presenting raw results of automatic coding on each dimension is not enough to indicate the collaborative process quality to teachers and students. Therefore, a new rating model that can assess and visualize the quality of collaborative process is proposed.**

***Keywords-CSCL; coding scheme; deep learning methods, automatic coding***

## I. INTRODUCTION

This article is an extended version of a conference paper presented at eLmL 2018, the Tenth International Conference on Mobile, Hybrid and On-line Learning [1]. It introduces more information on the related work of this study and especially a new proposal for visualization of results realized by our automatic coding method.

### *A. Analysis on Collaborative Process*

One of the greatest research topics in the actual Computer Supported Collaborative Learning (CSCL) research is to analyze its social and cognitive processes in detail in order to clarify what kinds of knowledge and meanings were shared within a group as well as how and by what arguments knowledge construction was performed. In addition, it is also required to develop CSCL system and tools with scaffolding function which may activate collaborative process by utilizing such knowledge.

However, because main data for the collaborative process analysis include contributions over chatting, images and voices on tools such as Skype, and various outputs prepared in the course of collaborative learning, it is totally inadequate to perform just quantitative analysis in order to analyze such data. Therefore, CSCL research changed direction more or less to qualitative research [2]-[5].

As these qualitative studies often result in in-depth case study, however, they have a downside that it is not easy at all to derive guidelines with generality, which are applicable also to other contexts. Therefore, studies have been conducted in recent years based on an approach of verbal analysis in which labeling for appropriately representing properties (hereinafter referred to as coding) is performed to each contribution in linguistic data of certain volume generated over the collaborative learning from perspectives of linguistics and collaborative learning activities [6]. On the other hand, an advantage of the approach is its capability of quantitative processing for significantly large scale data while keeping qualitative perspective. However, it is a task requiring significant time and labor to perform coding manually and it is expected to become impossible to perform coding manually in a case that data becomes further bigger in size.

In our research project, we have achieved certain results in a series of previous studies reported last year in eLmL 2017 and the like, using deep learning technique for automatic coding of vast amount of collaborative learning data [7][8[[9]. In this paper, while verification is performed for accuracy of the automatic coding based on deep learning technique similarly to last year, supervised data has been constructed by conducting coding manually depending on adopted multi-dimensional coding scheme in order to newly analyze collaborative learning process in a more multilateral and comprehensive manner.

### B. Purpose of This Study

The final goal of our research project is to implement support at authentic learning and educational settings such as real time monitoring of collaborative process and scaffolding for inactive groups based on analyses of large scale collaborative learning data as mentioned above.

As a further development of our previous research, a technique for automatizing coding of chat data is developed based on a multi-dimensional coding scheme capable of expressing collaborative learning process more comprehensively and its accuracy is verified in this study.

Specifically, after newly performing coding manually for substantial amount of the same chat data, which was used in the previous studies, a part of it is learned as training data by deep learning methods and then automatic coding is conducted for the test data. For accuracy verification, we try to verify the accuracy of automatic coding by calculating precision and recall of automatic coding of test data in each dimension. We also evaluate what type of misclassification occurred frequently in each dimension.

### C. Structure of This Paper

This paper is structured as follows. In Section II, we present the related work. The outline and results of our previous work are shown in Section III. Our coding scheme newly developed this time is described in Section IV. Section V presents the dataset with the statistics of the new coding labels assigned by the human coders. Our experiments and results of the study are shown in subsequent Section VI. Section VII proposes an assessment model of the quality of collaborative processes and envisage a possible visualization of this model. Finally, in Section VIII, we present the conclusion and future work to complete the paper.

## II. RELATED WORK

Deep neural networks [10] often has been applied in the field of natural language processing. Text classification is an important task in natural learning processing, for which various deep learning methods have been exploited extensively in recent studies. There are various modifications using convolutional neural networks (CNNs) that are applied for text classification [11][12][13]. In usual methods, texts are basically fed into CNNs with word-level embedding. Recent studies [14][15] show that character-level embedding is also promising method especially when datasets is sufficiently large. Using recurrent neural network (RNN) is another promising approach to achieve highly accurate results in text classification tasks. Long short-term memory units (LSTMs) [16] and gated recurrent units (GRUs) [17] are sophisticated architecture developed recently to overcome the drawbacks of RNNs. The language models used those RNNs can significantly outperform statistical language models, such as n-grams. RNNs are applied to text classification in various ways [17][18][19][20]. For instance, Yang et al. used a bidirectional GRU with attention modeling by setting two hierarchical layers that consist of the word and sentence encoders [21].

In the field of CSCL, some researchers have tried to apply text classification technology to chat logs. The most representative studies would be Rosé and her colleagues' works [22][23][24]. For example, they applied text classification technology to a relatively large CSCL corpus that had been coded by human coders using the coding scheme with multiple dimensions, developed by Weinbeger and Fisher [23][25]. McLaren's Argunaut project took a similar approach: he used online discussions coded manually to train machine-learning classifiers in order to predict the appearance of these discussions characteristics in the new e-discussion [26]. However, it should be pointed out that all these prior studies rely on the machine learning techniques before deep learning studies emerge.

## III. PREVIOUS WORK OF THIS STUDY

Outline of our previous work [7] is shown below.

### A. Conversation Dataset

Dataset for the study conducted last year is based on conversations among students participating in online collaborative learning. This data set is obtained from chat function of CSCL systeme originally developed by the authors for lectures in the university [27]. By the way, we will add that this data is also used in this paper. Usage situation of CSCL as the source of the dataset is shown in Table I. Since students participated in multiple classes, number of participant students is less than the number obtained by multiplying number of groups and that of group members.

TABLE I. CONTRIBUTIONS DATA USED IN THIS STUDY

| | |
|---|---|
| **Number of Lectures** | 7 Lectures |
| **Member of Groups** | 3-4 people |
| **Learning Time** | 45-90 mintutes |
| **Number of Groups** | 202 groups |
| **Number of Students** | 426 students |
| **Dataset** | 11504 contributions |

### B. Coding Scheme

According to a manual for coding prepared by the authors, a label was assigned to each contribution of chat. Any of the 16 types of labels as shown in Table II was assigned. The ratio of each label is shown in Figure 1.

### C. Automatic Coding Approach Based on Deep Learning

In the previous study, we adopted three types of Deep Neural Network (DNN) structures: 1) Convolutional Neural Networks (CNN), 2) Long-Short Term Memory (LSTM) and 3) Sequence to Sequence (Seq2Seq). Of the three models, Seq2Seq model is a deep neural network consisting of two LSTM units called encoder and decoder, and learning of classification problem and sentence generation is performed by entering pairs of strings of words to each part [28][29]. For example, the pair corresponds to a sentence in certain language and its translated sentence in case of translation

system as well as to question sentence and response sentence in case of question and answer system, respectively.

In addition, a model based on Support Vector Machine (SVM), which is a traditional machine learning approach is used as a baseline. Accuracy of each model is verified by comparing automatic coding concordance rate and Kappa coefficient. About technology and experiment results in detail for each classification model, please refer to existing literatures of the authors [7][8[[9].

TABLE II. LIST OF LABELS

| Label | Meaning of label | Contribution example |
|---|---|---|
| Agreement | Affirmative reply | I think that's good |
| Proposal | Conveying opinion, or yes/no question | How about five of us here make the submission? |
| Question | Other than yes/no question | What shall we do with the title? |
| Report | Reporting own status | I corrected the complicated one |
| Greeting | Greeting to other members | I'm looking forward to working with you |
| Reply | Other replies | It looks that way! |
| Outside comments | Contribution on matters other than assignment contents / Opinions on systems and such | My contribution is disappearing already; so fast! / A bug |
| Confirmation | Confirm the assignment and how to proceed | Would you like to submit it now? |
| Gratitude | Gratitude to other members | Thanks! |
| Request | Requesting somebody to do some task | Can either of you reply? |
| Correction | Correcting past contribution | Sorry, I meant children |
| Disagreement | Negative reply | I think 30 minute is too long |
| Complaint | Dissatisfactions towards assignments or systems | I must say the theme isn't great |
| Switchover | A contribution to change event being handled, such as moving on to the next assignment | Shall we give it a try? |
| Joke | Joke to other members | You should, like, learn it physically? : ) |
| Noise | Contribution that does not make sense | ?meet? day??? |

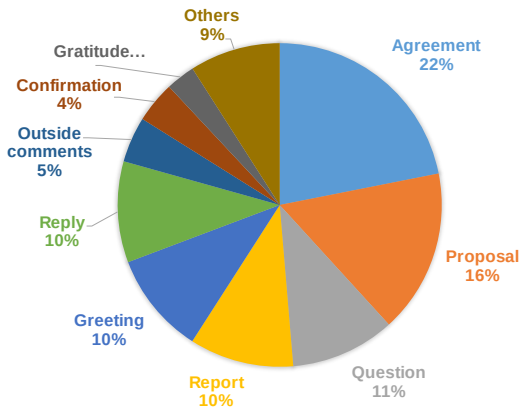


Figure 1. Ratio of each conversational coding labels

## D. Experiment and Assessment

### 1) Outline of experiment

For the data set with manually prepared coding labels as described above, we compared the prediction accuracy of automatic coding for each model.

With separation of sentences into morpheme using MeCab conducted at first as a preprocessing of data, words with low use frequency were substituted by "unknown". Subsequently, just 8,015 contributions were extracted and 90% and 10% of them were sorted into data for training and test, respectively.

Naive Bayes, Linear SVM, and SVM based on RBF Kernel were applied as baseline approaches.

### 2) Experiment Results

Table III shows prediction accuracy (concordance rate) of models proposed in the previous study and those adopted as baseline for test data. The concordance rate here refers to a proportion that manually assigned label conforms with predicted label output by a model. It is proved, as Table III shows, that accuracy of the proposed model's result is higher than that of baseline model. Among the three models as described above, it is found that there is almost no difference in concordance rate between the approaches based on CNN with word vectors trained using the Wikipedia data slightly enhanced accuracy and LSTM (0.67-0.68). These approaches show concordance rates a little bit higher (around 2 to 3%) compared with SMV as a baseline approach (0.64-0.66).

On the other hand, a model based on Seq2Seq showed the highest concordance rate among all of the models (0.718), higher by 5 to 7% and 3 to 4% compared with SVM and other models, respectively.

TABLE III. PREDICTIVE ACCURACIES FOR BASELINES AND DEEP-NEURAL-NETWORK MODELS

| Naïve Bayes | | SVM(Linear) | | SVM(RBF Kernel) | |
|---|---|---|---|---|---|
| unigram | uni+bigram | unigram | uni+bigram | unigram | uni+bigram |
| 0.554 | 0.598 | 0.642 | 0.659 | **0.664** | 0.659 |

| CNN | | LSTM | | Seq2Seq | |
|---|---|---|---|---|---|
| with wikipedia | w.o. wikipedia | single-direction | bidirection | bidirection | bidir. w. interm. |
| 0.686 | 0.677 | 0.676 | 0.678 | **0.718** | 0.717 |

Then, results as described above are discussed using Kappa coefficient, which is a measure of agreement between the two individuals (human and model in this case). At first, it may be said that LSTM model has achieved sufficiently higher result as the Kappa coefficient for the model shows 0.63. In general, Kappa coefficient of 0.8 or higher is believed to be preferable for utilizing automatic coding discrimination result by a machine in a reliable manner, however, further higher concordance rate is required. In case of Seq2Seq model, on the other hand, Kappa coefficient is 0.723 with great improvement, if not reaching 0.8.

The experiment results above have suggested that Seq2Seq model is superior to other approaches due to consideration for context information. Since Seq2Seq is a model with reply sources entered, it is believed that the improvement in the accuracy has been partly caused by not separate capturing of each contribution but consideration of the context information.

## IV. NEW CODING SCHEME

As our previous studies mentioned some cases that Replay may include a meaning of Agree in the coding scheme, the fact that the definition of one label may sometimes overlap the definition of another label has become a factor making it difficult to assign a label always with accuracy and reliability. In addition to these technical problems, more importantly, labels based on speech acts, which express the linguistic characteristics of the conversation are insufficient for the analysis of the learning process. With this single linguistic scheme, it is almost impossible to realize whether members of

a group engage in activities to solve the task, how members coordinate each other in terms of task division, time management, etc. during their collaboration, how each member constructs his argument, how members discuss and negotiate each other. From those described above, we propose a new coding scheme so that the automated coding accuracy will improve and that we may understand more accurately and globally collaborative process.

Our new coding scheme is constructed based on the multi-dimensional coding scheme proposed by Weinberger et Fischer [25]. As shown in Table IV, our scheme consists of five dimensions, while Weinberger and Fischer's one has four dimensions without Coordination dimension. We provide labels basically regarding a contribution as a unit similarly to way we used in the previous studies. In addition, while such values as number of contributions are provided as Participation dimension labels, those in other four dimensions are provided by selecting one label from among multiple labels. In other words, since one label is given for each dimension for one contribution, a plurality of labels will be assigned to one contribution. Therefore, the coding work with this scheme is extremely complicated and takes a lot of time, but the merit of automated coding is even greater. Each dimension is described in detail below.

TABLE IV. NEW CODING SCHEME

| Dimension | Description |
|---|---|
| Participation | Frequency of participation in argumentation |
| Epistemic | How to be directly involved in problem solving |
| Argumentation | Ideal assertion in argumentation |
| Social | How to cope with others' statements |
| Coordination | How to coordinate to advance discussion smoothly |

### A. Participation Dimension

Participation dimension is for measuring degree of participation in arguments. As this dimension is defined as quantitative data including mainly number of contributions and its letters, time of contributions, and interval of contributions, coding is performed by statistical processing on the database while requiring neither manual nor artificial intelligent coding. The list is shown in Table V.

TABLE V. PARTICIPATION DIMENSION

| Category | Description |
|---|---|
| Number of contributions | Number of contributions of each member during sessions |
| Number of letters of a contribution | Number of letters during a single speech |
| Time for contribution | Time used for a contribution |
| Interval of contributions | Time elapsed since last contribution |
| contributions distribution | Standard deviation of each member within a group |

Since Participation dimension labels handle number of specific contributions, it is possible to analyze quantitatively different aspects of participation in conversations but impossible to perform qualitative analysis such as whether the conversation contributed to problem solving.

### B. Epistemic Dimension

This dimension shows whether each contribution is directly associated with problem solving as a task and the labels are classified depending on contents of the contributions as shown in Table VI. This dimension's labels are assigned to all contributions.

TABLE VI. LABELS IN EPISTEMIC DIMENSION

| Label | Description |
|---|---|
| On Task | contributions directly related to problm solving |
| Off Task | contributions without any relationship with problem solving |
| No Sense | contributions with nonsensical contents |

Weinberger and Fischer's scheme has 6 categories to code epistemic activities, which consist in applying the theoretical concepts to case information. But, as shown in Table VII, we set only two categories here, because we want to give generality by which we can handle as many problem-solving types as possible. "On Task" here refers to contributions directly related to problem solving and such contributions with contents as shown below belong to "Off Trask".

- Contributions to ask meaning of problems and how to proceed with them
- Contributions to allocate different tasks to members
- Contributions regarding the system

Since Epistemic dimension represents whether directly related to problem solving, it works as the most basic code for qualitative analysis. In case of less "On Task" labels, for example, it is believed that almost no effort has been made for the task.

Besides, labels of Argument and Social dimensions are assigned when Epistemic dimension is "On Task", whereas those of Coordination dimension are assigned only when it is "Off Task". Coordination Dimension

Coordination dimension code is assigned only when Epistemic code is "Off Task" and it is also assigned to such contributions that relate to problem solving not directly but indirectly. A list of Coordination dimension labels is shown in Table VII but the labels are assigned not to all contributions of "Off Task" but just one label is assigned to such contributions that correspond to these labels. In addition, in case of replies to contributions with Coordination dimension labels assigned, labels of the same Coordination dimension are assigned.

"Task division" here refers to a contribution to decide who to work on which task requiring division of tasks for advancing problem solving. "Time management" is a contribution to coordinate degree of progress in problem solving, and for example, such contributions fall under the definition that "let's check it until 13 o'clock," and "how has it been in progress?" "Meta contribution" refers to a contribution for clarifying what the problem is when intention and meaning of the problem is not understood. "Technical coordination" refers to questions and opinions about how to use the CSCL System. "Proceedings" refer to contributions for coordinating the progress of the discussion.

Since Coordination dimension labels are assigned to such contributions that intend to problems smoothly, it is believed to be possible to predict progress in arguments by analyzing timing when the code was assigned. Further, in case of less labels of Coordination dimension, it may be predicted that smooth relationship has not been created within the group.

On the other hand, if a large number of these labels were assigned in many groups, it may be understood that there exists any defect in contents of the task or system.

TABLE VII. LABELS OF COORDINATION DIMENSION

| Label | Description |
|---|---|
| Task division | Splitting work among members |
| Time management | Check of temporal and degree of progress |
| Technical coordination | How to use the system, etc. |
| Proceedings | Coordinating the progress of the discussion. |

### C. Argument Dimension

Labels of Argument dimension are provided to all contributions, indicating attributes such as whether each contribution includes the speaker's opinion and whether the opinion is based on any ground. Labels of this dimension are provided to just one contribution content without considering whether any ground was described in other contribution.

A list of Argument dimension labels is shown in Table VIII. Here, presence/absence of grounds is determined whether any ground to support the opinion is presented or not but it does not matter whether the presented ground is reliable or not. A qualified claim represents whether it is asserted that presented opinion is applied to all or part of situations to be worked on as a task. "Non-Argumentative moves" refer to contributions without including any opinion and simple questions are also included in this tag. Also, as a logical consequence, this label is assigned to all off-task contribution in the Epistemic dimension.

TABLE VIII. LABELS IN ARGUMENT DIMENSION

| Label | Description |
|---|---|
| Simple Claim | Simple opinion without any ground |
| Qualified Claim | Opinion based on a limiting condition without any ground |
| Grounded Claim | Opinion based on grounds |
| Grounded and Qualified claim | Opinion with limitation based on grounds |
| Non-argumentative moves | contribution without containing opinion (including questions) |

Labels in Argument dimension are capable of analyzing the logical consistency of contribution contents. For example, if a contribution is filled just with "Simple Claim" it is assumed as a superficial argument.

In comparison with Weinberger and Fischer's scheme, we do not set for now the categories of macro-level dimension in which single arguments are arranged in a line of argumentation such as arguments, counterarguments, reply, for the reason that it seems difficult that the automatic coding by deep learning methods for this macro dimension works correctly. Social Dimension

Labels in Social dimension are provided when Epistemic code is "On task" but they are provided not to all contributions "On task" but to a contribution which conforms to Epistemic code. This dimension represents how each contribution is related to those of other members within the group. Therefore, it is required to understand not only a contribution but also the previous context. Table IX shows a list of labels of the dimension.

"Externalization" refers to contributions without reference to other's contributions and it is assigned to contributions to be an origin of arguments mainly at the start of argument on a topic. "Elicitation" is assigned to such contributions that request others for extracting information including question. "Consensus building" refers to contributions that express certain opinion in response to other's contribution and they are classified into the three labels below. "Quick consensus building" is assigned to such contributions that aim to form prompt consensus with other's opinion. It is assigned to a case to give consent without any specific opinion. "Integration-oriented consensus building" is assigned to such contributions that intend to form consensus with other's opinion while adding one's own opinion. "Conflict-oriented consensus building" is assigned to such contributions that confront with other's opinion or request revision of the opinion. "Summary" is assigned to contributions that list or quote contributions that have been posted.

Since Social dimension code represents involvement with others, it may be understood how actively the argument was developed or whose opinion within the group was respected by analyzing Social dimension labels. For example, it may be assumed that arguments with frequent "Quick consensus building" result in accepting all opinions provided with almost no deep discussion.

TABLE IX. CODE OF SOCIAL DIMENSION

| Label | Description |
|---|---|
| Externalization | No reference to other's opinion |
| Elicitation | Questionning the learning parner or proviking a reacion from the learning partner |
| Quick consensus building | Prompt consensus formation |
| Integration-oriented consensus building | Consensus formation in an integrated manner |
| Conflict-oriented consensus building | Consensus forming based on a confrontational stance |
| Summary | Statment listing or quoting contributions |

### D. Learning for each code granting and artificial intelligence

In the new coding scheme, "Participation" dimension labels are automatically generated from contribution logs, whereas other labels require manual coding by human coders in order to build up training data for deep learning and test data. Further, labels to be provided are decided by selecting from any of the dimensions of "Argument", "Social" and "Coordination" depending on a result of "Epistemic" labels. "Argument" and "Social" dimension labels are provided if the "Epistemic labels are "On task." In a case that "Epistemic" labels are "Off task", those in "Coordination" dimension are provided.

## V. DATASET AND STATISTICS

### A. Target Dataset

The raw dataset is taken from the real conversation log of the CSCL system, which is the same one as that of previous study (Table I). On this dataset, the coding labels were newly annotated based on the new coding scheme. Labeling was manually carried out by several people in parallel. The human coders were lectured about the new coding scheme by a professional in advance in order to code labels as accurately as possible. To evaluate the accuracy of the manual coding, we had each contribution annotated by two annotators and measured the coincidence rate for each dimension of the new coding scheme.

### B. Manual Coding and Preprocessing

While 9,962 contributions were manually coded in all, some contributions do not make sense as a text of CLSL. For instance, the duplicated posts, the blank posts, and the contributions that consist of only ASCII art can be mentioned. Such kinds of contributions were marked as "non-sense" when the annotators labeled, and removed or simplify ignored when the computer read them. After that, 9,197 contributions were remained as the useful data, on which the substantial jobs such as learning and classification are feasible.

The coincidence rates of the coding labels given by two human coders are significant for understanding the difficulty of the prediction, as well as to see the correctness of the manually coded labels. Table X shows the coincidence rate, the number of the valid contributions, and that of the coincidence contributions for each dimension. For the Epistemic dimension, since the coincidence rate is high for human coders, we can expect that it is also easy for machines to classify them. On the other hand, for the Social dimension, since the coincidence rate is low for human coders and the valid samples are sparse, the opposite result is expected.

TABLE X. THE VALID CONTRIBUTIONS AND THE COINCIDENT CONTRIBUTIONS

| | # of valid contributions | # of coincidence contributions | Rate |
|---|---|---|---|
| Epistemic | 9,197 | 8,460 | 0.92 |
| Argumentation | 9,083 | 7,765 | 0.85 |
| Coodination | 4,543 | 3,510 | 0.77 |
| Social | 3,917 | 2,619 | 0.67 |

### C. Statistics of the New Coding Labels

In this subsection, we describe the statistics of the new coding labels assigned by the human coders with respect to each dimension. As we have multiple coders classify them, the statistics depend on the coders. When making a dataset for machines, we limit the contributions so as to have the same label assigned by the human coders. Thus, we describe the statistics of such contributions.

The ratios of “On task” and “Off task” in the Epistemic dimension are shown in Figure 2. In our dataset, the ‘On task’ contributions were a bit fewer than the ‘Off task.’ This implies that, at least from the view point of the conversation log, the cost of the communication was more than the cost of discussion in group work. Although this result is just an instance obtained by applying our CSCL system to the actual group works for limited lectures, we can at least conclude that the communication cost is not small in a group work.

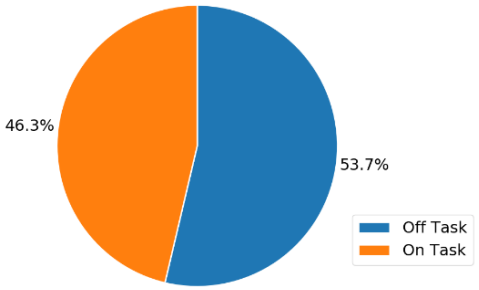


Figure 2. Ratio in the Epistemic dimension

Figure 3 shows the ratios of the labels in the Social dimension. Recall that its domain is On-task contributions. The label “Externalization” accounted half of the On-task contributions. The “Quick consensus building” followed it. Meanwhile, the ratios of the “Summary” and the “Consensus Buildings” except for the “Quick” one were small. These statistics show that the actual discussion mainly consisted of expressions of their opinions. Although we found that the contributions building consensus rarely come up in a real group work, we believe that they are the important keys for the discussion. Thus, we may can weight them when we assess the contribution to the discussion by students.

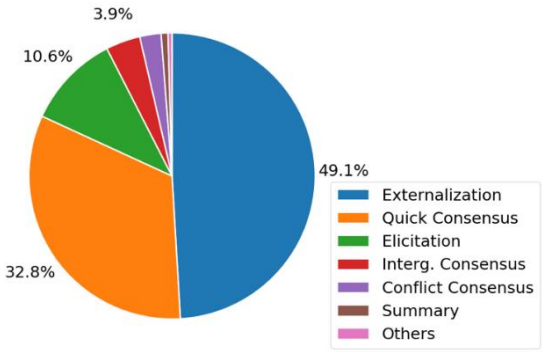


Figure 3. Ratio in the Social dimension

With respect to the "Coordination" dimension, the domain of which is the Off-task contributions, the most of them are assigned to "Other" as Figure 4 shows. The contributions labeled "Other" consist of short sentences that are not significant for neither discussion nor coordination of the group work. The representative examples are greetings and kidding. Meanwhile, the statistics show that the contributions except for "Other" also occupies more than a quarter. Since these kinds of contributions are related to coordinating tasks in the group work, they can be thought as important contributions for the assessment.

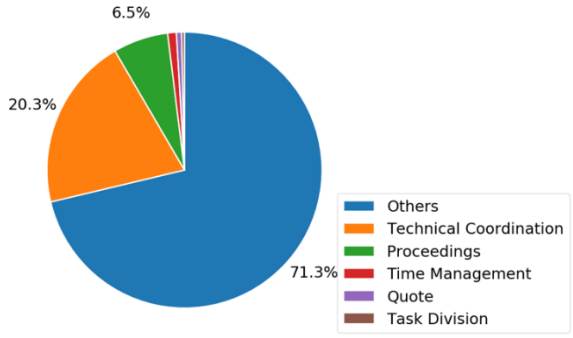


Figure 4. Ratio in the Coodination dimension

The labels in the "Argument" dimension are assigned independently of other dimensions. Thus, its domain spans both the On-task and the Off-task contributions. As shown in Figure 5, the label "Non-Argumentative moves" occupied more than 60 % of all. The label "Simple Claim" occupied the second percentage. To assess the discussion of the group work, at least it is necessary to remove the "Non-Argumentative" contributions and pay attention to which kind of claim is presented, even if almost every claim can be classified into the "Simple Claim". Therefore, the automatic coding for this dimension is as valuable as for the other three dimensions.

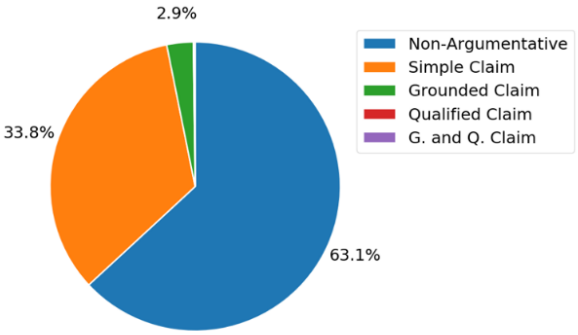


Figure 5. Ratio in the Argument dimension

## VI. EXPERIMENTS

### A. Approach to Learning and Classifcation

As described in Section II, deep neural networks (DNNs) outperform other machine learning methods significantly at least for the coding labels proposed by our previous studies [6][7][8]. Their results of the experiments show that the Seq2Seq-based model achieves the highest accuracy among several DNN structures. Thus, we apply the Seq2seq-based model to classify our new coding labels in this paper.

The new coding scheme has four axes to be labeled as discussed in Section III; the Epistemic, the Coordination, the Argument, and the Social dimension. In the following experiments, the labels in each axis, or the dimensions, are learned and classified. There are solid dependencies among the Epistemic, the Coordination and the Social dimensions, while the Argument dimension is independent of the other dimensions. As shown in Figure 6, there is a dependency tree among the former three dimensions. For instance, the label of the Social dimension is assigned only if that of the Epistemic is "On task." Therefore, the number of available contributions for learning is different for each classification task. In the following experiments, since we use the samples that have the coincidence labels only, the number of the available contribution was 8,460 for the Epistemic, 7,795 for the Augmentation, 3,510 for the Coordination, and 2,619 for the Social.

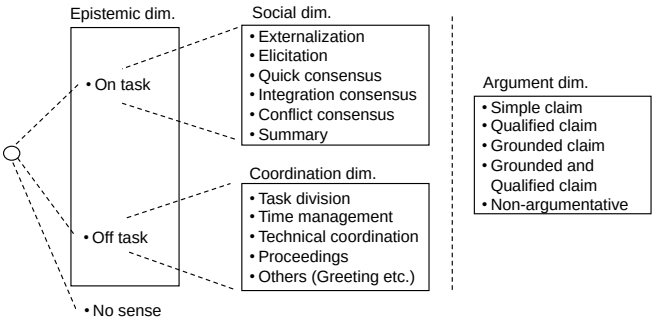


Figure 6. Dependency of Dimesions

### B. Parameter Settings

We set the parameters for learning to the same values as in our previous study. They include the various kinds of the parameters such as the number of layers, the vector sizes of layers, the option of the optimization algorithms, learning rate, etc. The details can be referred to our previous studies [6][7][8].

### C. Results for the Epistemic Dimension

The results of the experiments show that the On and Off tasks can be classified correctly with sufficiently high accuracy (Figure 7). The Seq2seq-based model achieves more than 90 % in both precision and recall (Table XI). Since the coincidence ratio by two human coders is 91%, we can say that the accuracy of automatic coding, which is comparable to human beings was obtained for the Epistemic dimension.

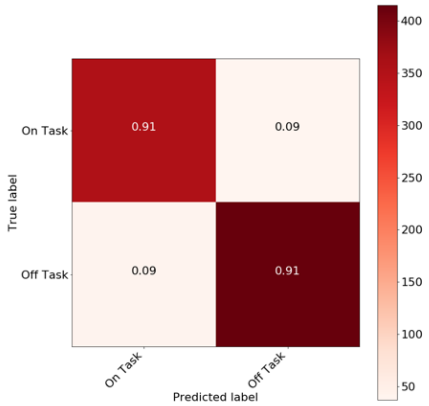


Figure 7. Confusion matrix for the Epstemic dimension

TABLE XI. PRECISION AND RECALL FOR THE EPSTEMIC DIMENTION

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| On Task | 0.90 | 0.91 | 0.90 | 390 |
| Off Task | 0.92 | 0.91 | 0.91 | 456 |
| Average(Micro) / Total | 0.91 | 0.91 | 0.91 | 846 |

### D. Results for the Argument Dimension

The classification accuracy is also high for the Argument dimension. The micro-averaged F1 score is 87 % (Table XII). Especially, the F1 score for the label "Non-argumentative Moves" is high sufficiently (92 %), which means that our model can surely recognize whether the contribution has any substantial meaning as a claim or not. On the other hand, while the precision for the "Simple Claim" is high (89 %), the recall for it is low (72 %). According to the confusion matrix shown in Figure 8, a quarter of the Simple Claim is misclassified into the Non-argumentative. This is because it is difficult to distinguish contributions that have a very small opinion from that have no opinions.

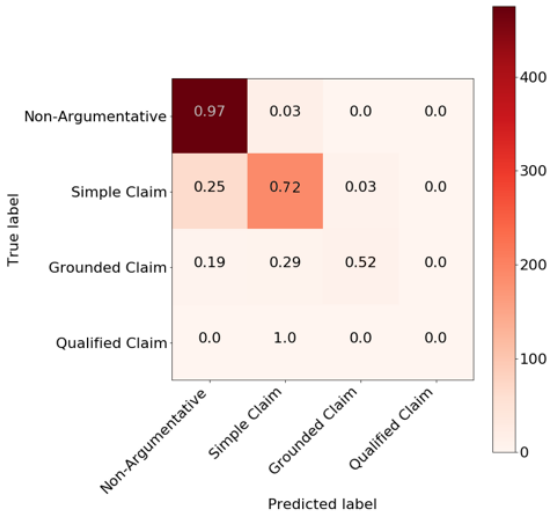


Figure 8. Confusion matrix for the Argument dimension

TABLE XII. PRECISION AND RECALL FOR THE ARGUMENT DEMENTIONFFIGU

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Non-Argumentative | 0.87 | 0.97 | 0.92 | 491 |
| Simple Claim | 0.89 | 0.72 | 0.80 | 264 |
| Grounded Claim | 0.58 | 0.52 | 0.55 | 21 |
| Qualified Claim | 0.00 | 0.00 | 0.00 | 1 |
| Average(Micro) / Total | 0.87 | 0.87 | 0.87 | 777 |

### E. Results for the Coordination Dimension

Regarding the Coordination dimension, our model also achieved high classification accuracy. Seeing that the number of supports varies greatly among the labels, we should evaluate the classification ability of the model by the micro-averaged accuracies over all coding labels. As Table XIII shows, the micro-averaged F1 score was 85 %.

According to the results for each label (Figure 9), the following is observed. The major labels such as "Other" and "Technical coordination" are classified correctly with high precisions, while the minor labels such as "Time Management", "Quote" and "Task Division" are not. Because the data for those minor labels are very limited, which have less than 50 contributions, it is quite difficult to learn them accurately. One of our future issues is to find some way to deal with those sparse labels.

TABLE XIII. PRECISION AND RECALL FOR THE COORDINATION DIMENTION

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Others | 0.91 | 0.91 | 0.91 | 242 |
| Technical Coordination | 0.81 | 0.80 | 0.81 | 82 |
| Proceedings | 0.58 | 0.70 | 0.64 | 20 |
| Time Management | 0.33 | 0.25 | 0.29 | 4 |
| Quote | 0.00 | 0.00 | 0.00 | 1 |
| Task Division | 0.00 | 0.00 | 0.00 | 2 |
| Average(Micro) / Total | 0.85 | 0.86 | 0.85 | 351 |

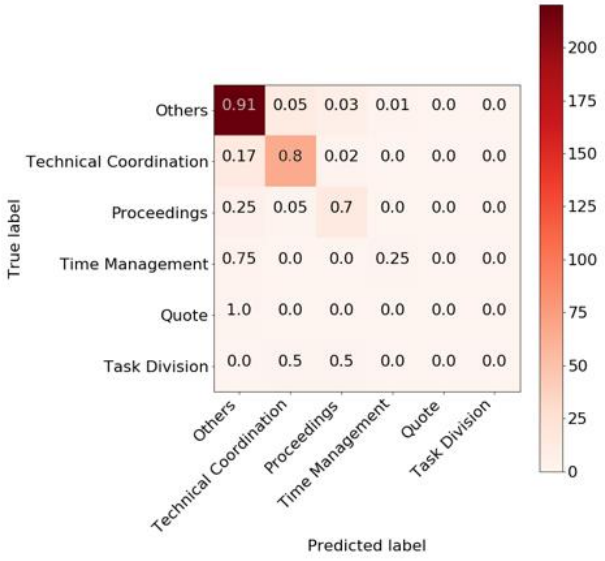


Figure 9. Confusion matrix for the Coodination dimension

### F. Results for the Social Dimension

Comparing to the other dimensions, the accuracy was relatively low for the Social dimension. The F1 score was 70 % (Table XIV). Since labeling the Social sometimes needs understanding the deep meaning of the contribution and the background story of the discussion, it seems to be difficult for machines to learn them correctly with limited data.

According to Figure 10, the recall of the label “Externalization” is especially low (61 %), while those of “Quick Consensus” and “Elicitation” are high sufficiently (93 % and 97 %, respectively). According to the confusion matrix in Figure 10, there is a major reason that worsen the accuracy; the Externalization labels are easily misclassified to the Quick Consensus and to the Elicitation, but not vice versa. This fact also explains the reason why the precisions for the Quick Consensus and the Elicitation are low though the recalls for them are high. To improve the result, it is necessary to pursue the causes of these two types.

TABLE XIV. PRESCISION AND RECALL FOR THE SOCIAL DIMENSION

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Externalization | 0.86 | 0.61 | 0.72 | 127 |
| Quick | 0.71 | 0.93 | 0.81 | 88 |
| Elicitation | 0.56 | 0.97 | 0.71 | 29 |
| Interg. Consensus | 0.17 | 0.14 | 0.15 | 7 |
| Conflict Consensus | 0.00 | 0.00 | 0.00 | 6 |
| Summary | 0.00 | 0.00 | 0.00 | 3 |
| Others | 0.00 | 0.00 | 0.00 | 2 |
| Average(Micro) / Total | 0.72 | 0.72 | 0.70 | 262 |

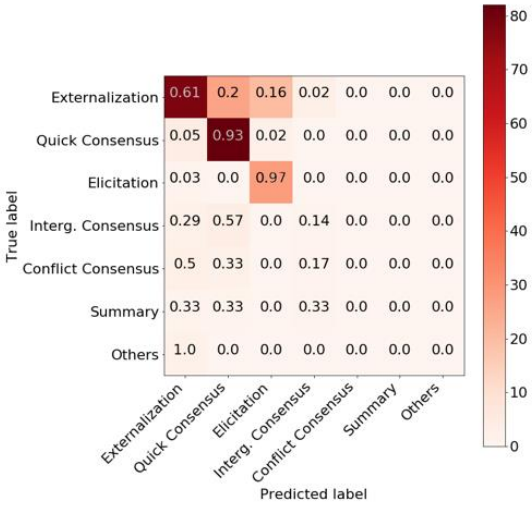


Figure 10. Confusion matrix for the Social dimension

## VII. ASSESSMENT MODEL

The method of automating multi-dimensional coding proposed in this study shows the technical possibility to clarify the situation of collaborative learning in real time from various perspectives even if it targets big educational data. However, as mentioned in the introduction, one of the ultimate goals of this research is to automatically analyze the collaborative learning process in real time and show the results to teachers and learners in an easy-to-understand manner. To that end, it is not enough to merely automate and speed up coding, and these coding data needs to be reintegrated and visualized in some form into a "readable" format.

So, we refer to the rating scheme for collaborative process assessment proposed by Meir, Spada, Rummel and we propose a model that adapts this scheme to the context of our research [30][31]. There are two reasons for choosing this scheme. First, in the empirical assessment, it is shown that positive findings exist for inter-rater reliability, consistency, and validity of this scheme. Second, as these authors have already recommended, this rating scheme is designed assuming that it will be customized according to various collaborative learning situations.

When designing the rating scheme, they define five aspects as factors of successful collaborative learning from the content analysis of empirical data and theoretical consideration based on the survey of the learning theory literature. That is, Communication, Joint information processing, Coordination, Interpersonal relationship, Motivation. In addition, as shown in Table XV, nine assessment dimensions are set for these five aspects. In these assessment dimensions, quantitative assessment is performed on a five point grade scale respectively.

TABLE XV. FIVE ASPECTS OF THE COLLABORATIVE PROCESS AND THERESULTING NINE DIMENSIONS OF MEIR, SPADA AND RUMMEL'S RATING SCHEME

| Process dimensions |
|---|
| A.Communication |
| 1) Sustaining mutual understanding |
| 2) Dialogue management |
| B. Joint information processing |
| 1) Information pooling |
| 2) Reaching consensus |
| C.Coordination |
| 1) Task divison |
| 2) Time management |
| 3) Technical coordination |
| D. Intrepersonal interaction |
| 1) Reciprocal interaction |
| E. Motivation |
| 1)Individual task orientation |

Since this paper is not a place to discuss the details of this original rating scheme, we will briefly describe the aspects and dimensions proposed in our research below. Regarding the aspects, it follows the original scheme. About the definition of "assessment targets" (we use this term to avoid confusion with dimensions of coding scheme), considering the fact that the major fields of our research are lectures at large university classrooms and the fact that there is no significant difference between students in knowledge level before class, some customizations are done. Also, in each assessment target, which coding data is referred to is also described. All of the nine assessment targets shown below are assumed to be quantified on a five-point grade scale, and the eight targets other than the last Individual task orientation can be easily visualized with an octagonal radar chart as shown in Fig.11.

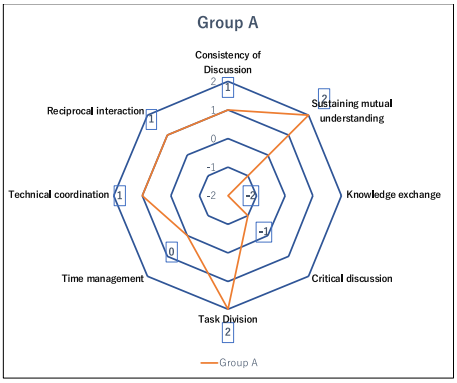


Figure 11. Example of visualization of collaborative process assessment

### A. Communication

In order to facilitate the discussion within the group, it is necessary that the basic concepts of the task and the hypotheses and problems that have appeared in discussions are shared. To that end, it is very important to repeatedly confirm that discussions are progressing on a common ground. Especially in conversation in chat, unlike face to face, members have to confirm mutual understanding more explicitly and frequently. We propose the following two as assessment targets of this smooth and "grounded" conversation.

*1) Consistency of Discussion*

For consistency of discussion, we evaluate whether discussions between members progress in a smooth flow. In a smoothly advancing discussion, other members should pay attention to others' contributions and return replay.

The reaction to the contribution by others is mainly expressed by the labels of Social dimension. To evaluate this target, it is good to refer to the following data and quantify it:

- Number of contributions in Social dimension and their frequency in On Task statements in Epistemic dimension;
- Number of contributions belonging to Social Dimension immediately after Social dimension's label "Externalization" and their frequency.

*2) Sustaining mutual understanding*

In maintaining mutual understanding, it is required to confirm in the process of discussion whether the understanding about basic concepts and problem awareness are shared between the members. Therefore, it is necessary to question each other, obtain answers, and confirm mutual understanding. Also, if there is a misunderstanding, it is necessary to resolve this.

For the assessment of this target, we will refer to the following data and quantify it:

- Number of contributions labeled as Elicitation in Social Dimension and frequency of contributions belonging to Social Dimension immediately after the Elicitation contribution.

However, the assessment of this target is insufficient in the current coding scheme and we will need to set a new coding label.

### B. Joint Information Processing

In problem solving for collaborative learning, it is required that each member provides his own knowledge, or obtains knowledge that he does not have from others to reach a more advanced solution. Therefore, the learner needs to explain knowledge in such a way that other members can understand. Also, in the process of problem solving, it is necessary to make questions and counterarguments to each other's opinions, and also to consider alternative solutions to make final decision making. The following two assessment targets are proposed as assessment targets of this aspect.

*1) Knowledge Exchange*

In this target, we assess to what extent members provided each other's knowledge in such a way that other members can understand their own knowledge.

For the assessment of this target, we will refer to the following data and quantify it:

- Number of contributions labeled as Grounded, Qualified, Grounded and Qualified in the Argument dimension and their frequency.

*2) Critical Discussion*

This target assesses if easy compromise is avoided and to what extent counterarguments and integrations to each other's views are effectuated.

For the assessment of this target, we will refer to the following data and quantify it:

- Number of contributions labeled as Conflict-oriented consensus building and Integration-oriented consensus building in the Social dimension and their frequency.

### C. Coordination

In order to succeed in collaborative learning, it is extremely important to plan ahead in advance of the whole work, to share subtasks with members, and to effectively manage time. In CSCL, not only chat but also interfaces such as file sharing and common workspace are generally prepared, so that technical coordination of these work environments also needs to be done successfully. As the assessment targets of this aspects, the following three assessment targets are adopted. In these three assessment targets, the small number of contributions is an indicator that coordination is insufficient. But if the number of contributions is too numerous, the problem solving activities (On task) may be not sufficiently carried out. Therefore, there is room for discussion about determining the appropriate number of contributions and their frequency.

*1) Task Division*

In this target, it is evaluated whether work division between members is done well.

For the assessment of this target, we will refer to the following data and quantify it:

- Number of contributions labeled as Task division in the Coordination dimension and their frequency.

*2) Time Management*

In this target, we evaluate how the members manage the time constraints and work.

For the assessment of this target, we will refer to the following data and quantify it:

- Number of contributions and frequency of Time management label in the Coordination dimension.

*3) Technical Coordination*

In this target, we evaluate whether effective use of technical resources is realized.

For the assessment of this target, we will refer to the following data and quantify it:

- Number of contributions labelled as Technical coordination in the Coordination dimension and their frequency.

### D. Interpersonal Relationship

In order for collaborative learning to succeed, it is necessary to exchange frankly opinions among participants; even if opinions are conflicting, it should be avoided that human relationship itself between members becomes confrontational. Participants must respect each other and behave in a friendly manner.

For the assessment of this aspect, the following assessment target is adopted. But it seems that it is insufficient to accurately capture this aspect in the current coding scheme. However, it is easy to automatically identify contributions such as greetings or thanks.

*1) Reciprocal Interaction*

In this target, we evaluate whether members are speaking equally or whether each participates evenly in problem solving and decision making.

For the evaluation of this target, we will refer to the following data and quantify it:

- Number of contribution and contribution distribution in Participation dimension.

### E. Motivation

In order to animate the group as a whole, it is necessary for each member to act positively on problem solving and encourage other members to actively participate. However, there are significant differences between members such as efforts to solve problems, participation in discussions, and encouragement to other group members. In order to evaluate the aspect of this individual contribution, the following assessment target is adopted.

*1) Individual Task Orientation*

In this target, the contribution degree of each group member is individually assessed.

For the assessment of this target, we will refer to the following data and quantify it:

- Number of contributions labelled as On task in the Epistemic dimension of each member.

## VIII. CONCLUSION AND FUTURE WORK

### A. Conclusion

In this study, we proposed a newly designed coding scheme with which we tried to automate time-consuming coding task by using deep learning technology.

We have constructed a new coding scheme with five dimensions to analyze different aspects of the collaboration process. After manually coding a large volume dataset, we proceeded to the machine learning of this dataset using Seq2seq model. Then, we evaluated the accuracy of this automatic coding in each dimension. Except some typical types of the misclassifications, the results were overall positive. If this misclassification is resolved to a considerable extent, it will also come into view to apply this technique in real educational settings and for large classes with many students in order to perform real-time monitoring of learning process or ex-post analysis of big educational data.

Finally, at the end of the paper, we propose a new assessment model that can assess and visualize the quality of collaborative process.

### B. Future Work

As for the future research directions, we may have three areas to pursue.

The first area is about some typical misclassifications in the Social Dimension. To improve prediction accuracy, one could make more explicit and comprehensible the referential relation between a contribution and others even for the machines, if one indicates contributions to which a contribution refers. For example, with regard to the typical misclassification mentioned above between “Externalization” and “Quick Consensus” or “Elicitation”, since contributions labeled “Externalization” have no reference to other contributions, we can hope to effectively reduce these misclassifications with this kind of indicator. In addition, as the next step of this paper, it seems to be worth trying to compare the accuracy using DNN models other than Seq2seq and other network structures such as memory networks [32].

The second area concerns the intrinsic structure of our coding scheme. Since the scheme contains different dimensions and under each dimension different labels are hierarchically organized, it is very interesting to discover not only correlations among dimensions, but also among labels belonging to different dimensions [33]. If we can input the information about the correlation between such labels in some form at the time of automatic classification, the accuracy of automatic coding can be further improved.

The third area relates to the assessment method and its visualization of collaborative process. In this paper, the method of calculating the rating of assessment targets is not defined yet, which is an urgent task. Furthermore, we will have to reconsider which data should be referenced for each target. Also, it may be necessary to partially modify the scheme itself to fit the assessment model. For visualization, we should consider not only visualization of real-time collaborative situation but also design method to intuitively visualize transition on time axis and comparison between different groups.

## ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 26350289, 17H02004 and 16K01134.

## REFERENCES

[1] T.Kanayama, Ch.Shibata, K.Ando, and T.Inaba, “Using deep learning methods to automate collaborative learning process coding based on multi- dimensional coding scheme,” The Tennth International Conference on Mobile, Hybrid, and On-line Learning, pp. 45-53, 2018.

[2] G. Stahl, T. Koschmann, and D. Suthers, "Computer-supported collaborative learning," In The Cambridge handbook of the learning science, K. Sawyer, Eds. Cambridge university press, pp. 479-500, 2014.

[3] P. Dillenbourg, P. Baker, A. Blaye, and C. O'Malley, "The evolution of research on collaborative learning," In Learning in humans and machines: Towards an interdisciplinary learning science, P. Reimann and H. Spada, Eds. Oxford: Elservier, pp. 189-211, 1996.

[4] T. Koschmann, "Understanding understanding in action," Journal of Pragmatics, 43, pp. 435-437, 2011.

[5] T. Koschmann, G. Stahl, and A. Zemel, "The video analyst's manifesto (or The implications of Garfinkel's policies for the development of a program of video analysis research within the learning science)," In Video reseach in the learning sciences, R. Goldman, R. Pea, B. Barron and S. Derry, Eds. Routledge, pp. 133-144, 2007.

[6] M. Chi, "Quantifying qualitative analyses of verbal data : A pratical gauide ," Journal of the Learning Science, 6(3), pp. 271-315, 1997.

[7] C. Shibata, K. Ando, and T. Inaba, "Towards automatic coding of collaborative learning data with deep learning technology", The Ninth International Conference on Mobile, Hybrid, and On-line Learning, 2017, pp. 65-71.

[8] K. Ando, C. Shibata, and T. Inaba, "Analysis of collaborative learning processes by automatic coding using deep learning technology", Computer & Education, 43, pp. 79-84, 2017.

[9] K. Ando, C. Shibata, and T. Inaba, "Coding collaborative learning data automatically with deep learning methods", JSi SE Research Report, 32, 2017.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, 521(7553), pp. 436-444, 2015.

[11] Y. Kim, "Convolutional neural networks for sentence classification," arXiv preprint arXiv:1408.5882, 2014.

[12] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu, "Text classification improved by integrating bidirectional lstm with two-dimensional max pooling". In *Proceedings of COL ING 2016*, 2016.

[13] R. Johnson and T. Zhang, "Deep pyramid convolutional neural networks for text categorization", In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol- ume 1: Long Papers)*. volume 1, pp. 562–570 2017.

[14] X. Zhang, J. Zhao, and Y. LeCun. "Character-level convolutional networks for text classification," In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS2015), pp. 649-657, 2015.

[15] A. Conneau, H. Schwenk, Y. LeCun and L. Barrout, "Very Deep Convolutional Networks for Text Classification," In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 1107–111, 2017

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, 9(8), pp. 1735-1780, 1997.

[17] J. Chung, C. Gulcehre, K. Hyun Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," arXiv preprint arXiv:1412.3555, 2014.

[18] Z. Yang et al., "Hierarchical Attention Networks for Document Classification," In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics(NAACL2016), Human Language Technologies, 2016.

[19] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing(EMNLP2016), pp. 1422–1432, 2015.

[20] R. Johnson and T. Zhang, Supervised and semi-supervised text categorization using lstm for region embeddings. In *International Conference on Machine Learning*, pp. 526–534, 2016.

[21] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification", In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 328–339, 2018.

[22] C. Rosé et al., "Towards an interactive assessment framework for engineering design project based learning," In Proceedings of DETC2007, 2007.

[23] C. Rosé et al., "Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning," International Journal of Computer Supported Collaborative Learning, 3(3), pp. 237-271, 2008.

[24] G. Gweon, S. Soojin, J. Lee, S. Finger and C.Rosé, "A framework for assessment of student project groups on-line and off-line," In Analyzing Interactions in CSCL: Methods, Approaches and Issues, S. Putambekar, G.Erkens and C. Hmelo-Silver Eds. Springer, pp. 293-317, 2011.

[25] A. Weinberger and F. Fischer, "A frame work to analyze arugmetative knowledge construciton in computer-supported learning," Computer & Education, 46(1), pp. 71-95, 2006.

[26] B. McLaren, O. Scheuer, M. De Laat, H. Hever and R. De Groot, "Using machine learning techniques to analysze and support mediation of student e-discussions," In Proceedings of artificial intelligence in education, 2007.

[27] T. Inaba and K. Ando, "Development and evaluation of CSCL system for large classrooms using question-posing script," International Journal on Advances in Software, 7(3&4),pp. 590-600, 2014.

[28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv, pp. 1409.0473, 2014.

[29] O. Vinyals and Q. V. Le, " A Neural Conversational Mode," arXiv preprint arXiv:1506.05869, (ICML Deep Learning Workshop 2015), 2015.

[30] A. Meir, H. Spada, and N. Rummel, "A rating scheme for assessing the qualituy of computer-supported collaboraiton processes," International Journal of Computer-Supported Collaborative Learning, 2(1), pp. 63-86, 2007.

[31] N. Rummel, A. Deiglmayr, H. Spada, G. Kahrimanis, and N.Avouris, "Analyzing collaborative interactions across domains and settings: an adaptable rating scheme," in Analyzing interactions in CSCL, S.Puntambekar et al, Eds. Springer, pp. 367-390, 2011.

[32] S. Sukhbaatar, A. Szlam, J. Weston and R. Fergus, "End-to-end Memory Networks," Proceedings of the 28th International Conference on Neural Information Processing Systems, pp. 2440-2448, 2015.

[33] F. Scafino, G. Pio, M. Ceci, and D. Moro, "Hierarichical multi-dimensional classification of web documents with MultiWebClass," International Conference on Discovery Science, pp. 236-250, 2015.