International Journal of Smart Computing and Artificial Intelligence
International Institute of Applied Informatics
2024, Vol. 8, No. 2, IJSCAI837

Classroom Utterance Analysis and Visualization Using a Generative Deep Neural Networks for Dialogue Model

Sakuei Onishi *, Tomohiko Yasumori †, Hiromitsu Shiina ‡

Abstract

In elementary school and other classes at different levels, teachers have less time for reflection, which can involve self-assessment and classroom observation. Never-theless, reflection activities are becoming increasingly important. Unfortunately, few studies support teachers ' reflection, and studies that analyze classroom utterances treat them as text and do not consider them as dialogues. However, in the field of dialogue response generation, some dialogue models using neural networks have been proposed. In this study, we propose a dialogue model that considers the domain of the class and extracts similar utterances. The proposed model considers the domain of ele-mentary school classes, a domain in which speakers can be classified, and incorporates a method to abstract the characteristics of speakers by clustering. The proposed model can be constructed with a relatively small number of parameters. We also developed a system to visualize the classification probabilities analyzed using the proposed dia-logue model. The developed visualization system was evaluated by experts and was found to visually recognize the classification bias of an utterance and to confirm the quality of the utterance by the similarity of the utterances.

Keywords: Dialogue Model, Pedagogical Deep Learning, Generative Deep Neural Network, Classroom Utterance Analysis, Global Variational Transformer (GVT)

1 Introduction

The Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) Central Educational Council [1] lists the practical leadership qualities and abilities that teachers must have and emphasizes the importance of establishing effective teacher characteristics, which involve continuous exploration and learning. The MEXT 2017 elementary school guidelines [2] stated that classes should aim to realize "proactive and interactive deep learning." A wider variety of initiatives have been taken in elementary school classrooms, and the utterances of teachers and children have also been analyzed in empirical pedagogical studies. Of these, studies of reflection activities [3][4] are attracting considerable attention. From the perspective of proficiency research, it is said that teachers need to be adaptively proficient in developing their students into independent learners. In research on reflection based on pedagogical content knowledge (PCK) [5], adaptively proficient teachers demonstrated reflection on the basis of two aspects: instructor-centered

* Graduate School of Informatics, Okayama University of Science, Okayama, Japan
† A Faculty of Education, Okayama University of Science, Okayama, Japan
‡ A Faculty of Informatics, Okayama University of Science, Okayama, Japan

and learner-centered PCK, where learner-centered PCK was found to be an essential element along with instructor-centered PCK. A reflection matrix [5] has also been proposed that emphasizes collaborative learning along with the instructor and learner-centered PCK. From the perspective of "proactive and interactive deep learning," there has been extensive research focused on the digitization of the reflection method. Moreover, utterance analysis has been attempted using computer systems [6]. These studies have been undertaken because elementary and junior-high school teachers in Japan are extremely busy and survey results [7] show that the time spent on research classes and training on reflection in Japan is the shortest among 48 countries. For such busy teachers, the development of a machine-based analysis method allowing a teacher to engage in effective reflection without spending a significant amount of time is a pressing issue.

In actual elementary school classes, teachers conduct their classes while monitoring the children's situation, and it is considered rare for a teacher to conduct the class in a unilateral manner. Typically, the children have many opportunities to discuss their learning, express their opinions, and exchange impressions about the class with their classmates. Under these conditions, children are considered to be interacting in class. In addition, the introduction of flipped classrooms has also been contemplated in universities. Flipped classrooms are considered important for deepening interactions between students and teachers. As a certain degree of interaction is established between teacher and children, and between the children themselves, it is possible to automatically analyze utterances and dialogues that promote and demonstrate understanding of the class, which may offer a lot of feedback to teachers.

In natural language processing via machine learning, it is possible to process languages while considering context, for example by using models such as a bidirectional encoder representations from transformers (BERT) [8], which uses a transformer [9]. We have also seen research [10][11][12] on dialogue response generation for chatbots. Furthermore, a global variational transformer (GVT) [13] using a transformer has been proposed. In this study, in addition to proposing an extended GVTSC model that extracts the characteristics of the speaker using clustering in advance, The extended GVTSC model is a small model with a relatively small number of parameters and fits one-to-many dialogue. We analyze utterances using an extended GVTSC model related to interaction within the elementary school classroom. Specifically, we record elementary school mathematics classes and analyze the transcribed interactive-style text information. In this analysis, the annotation is manually attached to the utterances spoken during the class in advance. The annotation involves attaching a label that describes the learning type for utterances related to "proactive and interactive deep learning." After this annotation, utterances that are close to the target utterance are extracted using the proposed extended GVTSC model. We previously reported on class analysis by extracting similar utterances to class utterances using the extended GVTSC[14]. In this paper, we also propose a system for estimating labels for class utterances and visualizing class trends using the estimated labels. By using labeled utterances similar to class utterances, it is possible to estimate the labels of class utterances. However, using the estimated labels related to the utterances as an aid to reflect simply by looking at them is difficult. Therefore, the system must be developed such that the distribution of labels is visualized. This, in turn, will enable the utterances to be checked with video clips of the classes in chronological order.

Table 1: Proactive, interactive and deep learning labels

Broad classification	Narrow classification
Proactive learning	Provides perspective Proactive Run counter to children ' s proactive learning There are issues with the reflection. Teacher presents unilaterally. Purpose is to elicit the right answer
Interactive learning	Enables interaction Encourages interaction
Deep learning	Deep learning Functions Mathematical perspective

2 Used Data

2.1 Categories of Proactive, Interactive, and Deep Learning

Studies [15] defining "proactive and interactive deep learning" and arranging specific examples of utterances are broadly classified into "proactive learning," "interactive learning," and " deep learning," which shows the existence of utterances that overlap multiple categories. The ten types of categories for "proactive and interactive deep learning" used in this study are shown in Table 1. In terms of the definition of "proactive, interactive, and deep learning," there are generalized definitions common to each subject and field and subject-specific definitions. The target of this study is elementary school mathematics classes, consequently, we considered subject-specific classes, e.g., "mathematical perspective."

2.2 Classroom Dialogue Data

The classroom dialogue data are the interactive-type text information created after recording the elementary school mathematics class and transcribing the utterances of the teacher and the students. In this study, the annotations were manually applied to the utterances. Annotations were applied to utterances related to "proactive and interactive deep learning." A list of labels identifying the learning types is shown in Table 1.

The classroom data were obtained from a fourth grade mathematics (proportions) class and a sixth grade mathematics (proportions) class. In this paper, these two classes are referred to as class 1 and class 2, respectively. The class 1 data included 193 utterances, and 13 utterances were annotated. The class 2 data included 274 utterances, and 27 utterances were annotated.

3 GVT Model with Information of Each Speaker

The GVT model is one in which the CVAE model [16] and an RNN model [17][18][19] used for dialogue, has been rewritten as Transformer. For the GVT, the utterances are input as a context without differentiating the speakers. Thus, a latent variable is generated from the entire context, and the characteristics of each speaker are diluted. Therefore, we propose an extended GVT model in which utterances are separated and input for each speaker, and

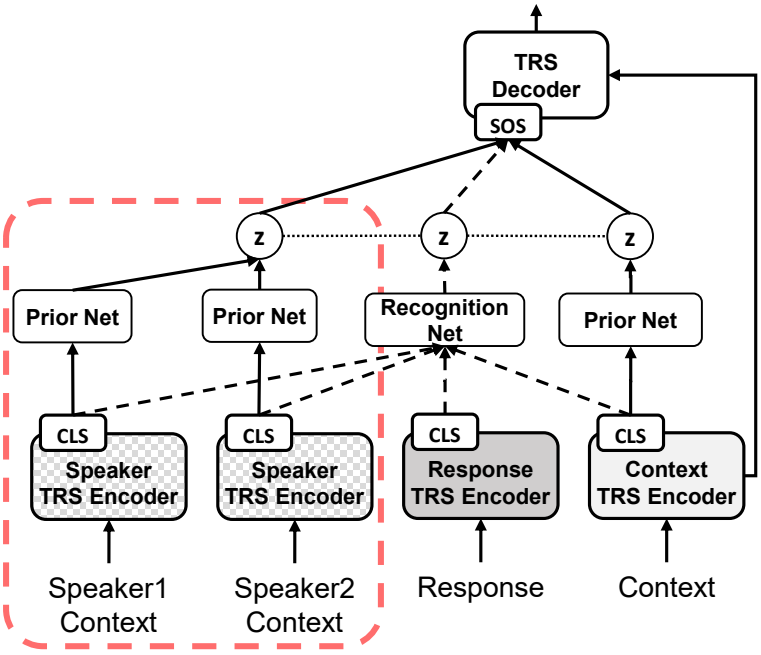

Figure 1: Structure of the extended GVT model

the latent variable of each speaker is utilized to consider the characteristics of each speaker. For the extended GVT, the structure considers changes in speakers and previous utterances. Figure 1 shows the extended GVT model. In contrast to the conventional GVT model, the proposed model differentiates utterances by each speaker and takes the context (dotted line in Figure 1) as input, where latent variable z is sampled from Prior Net for each speaker.

For the TRS encoder, the CLS token is added to the beginning of the input sequence, where the Transformer model calculates the output vector. The entire dialogue context is input to the context TRS encoder to obtain the output vector. The context summarizes the utterances of two people in the given dialogue, and each speaker can be identified and separated. To obtain the output vector, each speaker divides the context and inputs it to each speaker TRS encoder.

Prior Net and Recognition Net are used to sample z, where prior and posterior probability distributions are approximated using a multilayer perceptron (MLP). The Prior Net estimates the mean and variance of the context vector using the MLP based on the output vector of the speaker TRS encoder or the context TRS encoder CLS token. Note that z is sampled from the normal distribution that follows the mean and variance. For the Recognition Net, in addition to the speaker TRS encoder and context TRS encoder, the output vector of the response TRS encoder CLS token is also used to estimate the mean and variance of the vector of the entire dialogue using the MLP. Here, the latent variable z is sampled from a normal distribution, with the mean and variance estimated in the same manner as the Prior Net. The output vector of the TRS encoder CLS token can be considered a vector that expresses the entire input; thus, prior and posterior probability distributions are generated from the output vector of the CLS token to sample the latent variable z.

For the TRS decoder, latent variables are used to generate responses by inserting the latent variable of the response speaker and the normal latent variable of the SOS token at the beginning of the input sequence. In addition, the TRS decoder uses the latent variable

sampled by the Recognition Net while learning, and it uses the latent variable sampled from the Prior Net during generation.

The extended GVT model optimizes the model by maximizing the evidence lower bound (ELBO) as follows.

$$\begin{aligned}
&\mathscr{L}_{ELBO}(x,c) \\
&= \mathbb{E}_q(z|x,c)[\log p(x|z,c)] \\
&\quad - KL(q(z|x,c)\|p(z|c)) \\
&\quad - KL(s(z|x,c_{s1},c)\|r(z|c_{s1})) \\
&\quad - KL(s'(z|x,c_{s2},c)\|r'(z|c_{s2})) \\
&\leq \log p(x|c).
\end{aligned} \tag{1}$$

Here, c denotes the context, c_{s1} denotes the speaker 1 context, c_{s2} denotes the speaker 2 context, x denotes the response, and z denotes the latent variable. In addition, KL represents the Kullback-Leibler divergence between distributions and the prior probability distribution, and p, r, and r' can be defined as follows:

$$p(z|c) \sim \mathscr{N}(\mu_p, \sigma_p^2), \tag{2}$$
$$r(z|c_{s1}) \sim \mathscr{N}(\mu_r, \sigma_r^2), \tag{3}$$
$$r'(z|c_{s2}) \sim \mathscr{N}(\mu_{r'}, \sigma_{r'}^2), \tag{4}$$

where

$$\left[\mu_p, \log(\sigma_p^2)\right] = \mathrm{MLP}_p(c), \tag{5}$$
$$\left[\mu_r, \log(\sigma_r^2)\right] = \mathrm{MLP}_r(c_{s1}), \tag{6}$$
$$\left[\mu_{r'}, \log(\sigma_{r'}^2)\right] = \mathrm{MLP}_{r'}(c_{s2}). \tag{7}$$

The posterior probability distribution, q, s, and s' are defined as follows:

$$q(z|x,c) \sim \mathscr{N}(\mu_q, \sigma_q^2), \tag{8}$$
$$s(z|x,c_{s1},c) \sim \mathscr{N}(\mu_s, \sigma_s^2), \tag{9}$$
$$s'(z|x,c_{s2},c) \sim \mathscr{N}(\mu_{s'}, \sigma_{s'}^2), \tag{10}$$

where

$$\left[\mu_q, \log(\sigma_q^2)\right] = \mathrm{MLP}_q(x,c), \tag{11}$$
$$\left[\mu_s, \log(\sigma_s^2)\right] = \mathrm{MLP}_s(x,c_{s1},c), \tag{12}$$
$$\left[\mu_{s'}, \log(\sigma_{s'}^2)\right] = \mathrm{MLP}_{s'}(x,c_{s2},c). \tag{13}$$

KL annealing [20] and bag-of-words (BoW) loss [16][21] are incorporated as a result of KL vanishing, which occurs when the decoder stops considering the information of the latent variable z as the learning process progresses. KL annealing is a method in which the KL divergence value in Equation 1 is assigned a weight that increases linearly from 0 to 1 as learning progresses. The BoW loss method includes subtasks that estimate the set of words included in the response to strengthen the relationship between the latent variable and the words in the response.

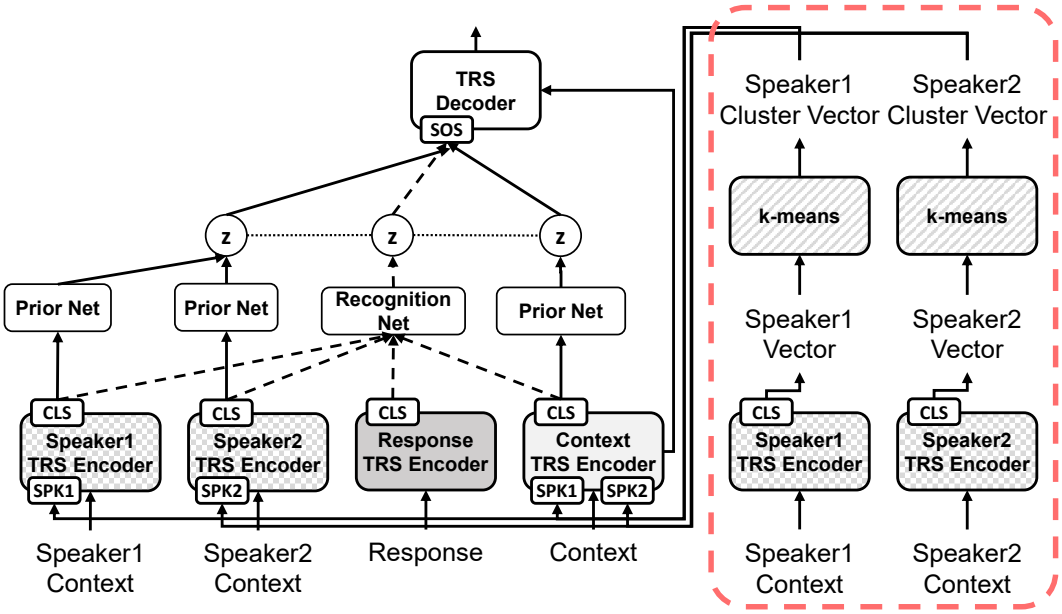

Figure 2: Structure of proposed extended GVTSC model

4 Extended GVTSC Model with Speaker Clustering

4.1 Overview

When generating dialogue responses, safe responses to various dialogues can be generated; however, the diversity of the responses may decrease [22][23]. The GVT model uses sampled latent variables as input to a decoder. Here, it is assumed that diverse responses can be achieved by expressing and sampling the characteristics of the speaker with latent variables. However, previous studies have shown that this tends to reduce the consistency of responses generated by the latent variables [24]. Thus, the characteristics of each speaker are abstracted using a clustering method, and an encoder considers the characteristics of the speaker to improve consistency and diversity. In the proposed extended GVTSC model, the speakers are clustered in advance.

4.2 Creating Speaker Characteristics through Clustering

The structure of the proposed extended GVTSC model is shown in Figure 2. In the extended GVT model, an encoder is implemented in the GVT model for each speaker. To realize this, clustering is employed to create a feature vector for the speaker (the dotted line in Figure 2), and this is used in the context encoding process.

The process flow for clustering to create speaker feature vectors is shown in Figure 3. The context is a summary of the utterances of the two interacting parties, and this can be analyzed for each speaker. Thus, the dialogue context is divided between speakers, and processing occurs for each speaker. Here, processing is the same for each speaker. First, the context for each speaker is encoded using the Speaker TRS Encoder. With this TRS Encoder, a CLS token is added to the start of the input sequence, and the output vector is calculated using the transformer. A CLS token vector is obtained as the context vector for each speaker (Speaker Vector in Figure 3). Next, k-means clustering is performed on

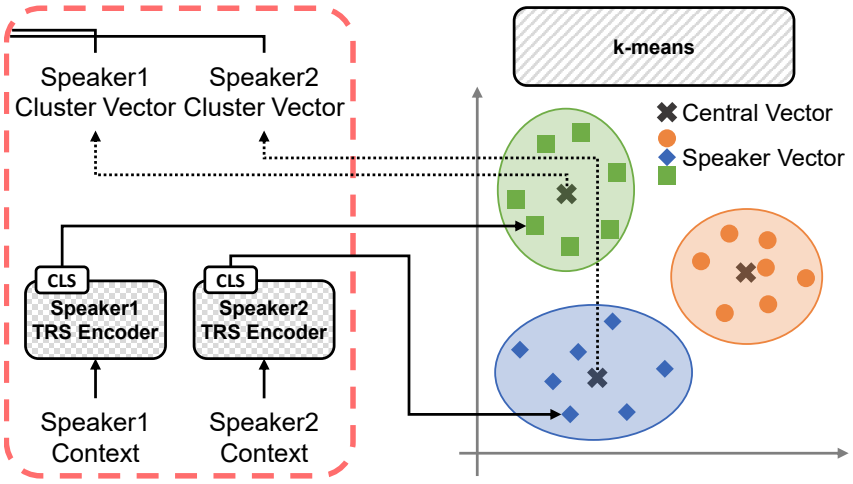

Figure 3: Process for clustering to create speaker feature vectors

the Speaker Vector. Note that the number of clusters k and the hyperparameters must be determined experimentally. As a result of clustering, the clusters to which the Speaker Vectors belong are predicted. Finally, the Central Vectors for these clusters (Speaker Cluster Vector in Figure 3) are obtained. Here, the TRS encoder used in the clustering is the same TRS encoder trained for response generation; however, backpropagation is not used when training the clustering process.

4.3 Dialogue Response Generation

The entire context of the dialogue is input to the context TRS encoder to obtain the output vector. With the context encoding, a token (SPK1, SPK2) is added for each speaker to the input sequence. Here, the Speaker 1 Cluster Vector is input to SPK1, and the Speaker 2 Cluster Vector is input to SPK2. In addition, the context is divided for each speaker, and each context is input into the respective speaker TRS encoder to obtain the output vector. Furthermore, the Speaker 1 token (SPK1) is added to the input sequence in speaker 1 TRS encoder and the Speaker 1 Cluster Vector, and the Speaker 2 token (SPK2) is added to the input sequence in speaker 2 TRS encoder and the Speaker 2 Cluster Vector. As the encoding for the context of each speaker utilized the feature vector of each speaker, the encoding attempts to consider the speaker ' s characteristics.

Sampling the latent variable z from the Prior Net and Recognition Net is approximated using MLPs for the prior and posterior distributions. Based on the output vector for the CLS token of either the speaker TRS encoder or context TRS encoder, the Prior Net estimates the mean and variance of the context vector using the MLP. Here, the latent variable z is sampled from the normal distribution following this mean and variance. With the Recognition Net, in addition to the speaker TRS encoder and context TRS encoder, the output vector of the response TRS encoder CLS token is used to estimate the mean and variance of the entire dialogue vector using the MLP. As with the Prior Net, the latent variable z is sampled from the normal distribution following this estimated mean and variance. As the output vector for the TRS encoder CLS token can be considered a vector representing the entire input, the prior and posterior distributions are generated from the output vector of the CLS token, and the latent variable z is sampled.

In the TRS decoder, the latent variable of the response speaker is input in addition to the normal latent variable to the SOS token at the beginning of the input sequence, and the latent variable is used to generate the response. In this case, the TRS decoder uses the latent variable sampled from the Recognition Net during training and the latent variable sampled from the Prior Net during generation.

4.4 Formulation of Extended GVTSC Model

The extended GVTSC model optimizes the model in the same manner as the extended GVT model by maximizing the ELBO as follows.

$$\begin{aligned}
&\mathscr{L}_{ELBO}(x,c) \\
&= \mathbb{E}_q(z|x,c,v_{\mathrm{s1}},v_{\mathrm{s2}})[\log p(x|z,c,v_{\mathrm{s1}},v_{\mathrm{s2}})] \\
&\quad - KL(q(z|x,c,v_{\mathrm{s1}},v_{\mathrm{s2}})\|p(z|c,v_{\mathrm{s1}},v_{\mathrm{s2}})) \\
&\quad - KL(s(z|x,c_{\mathrm{s1}},c,v_{\mathrm{s1}},v_{\mathrm{s2}})\|r(z|c_{\mathrm{s1}},v_{\mathrm{s1}})) \\
&\quad - KL(s'(z|x,c_{\mathrm{s2}},c,v_{\mathrm{s1}},v_{\mathrm{s2}})\|r'(z|c_{\mathrm{s2}},v_{\mathrm{s2}})) \\
&\leq \log p(x|c),
\end{aligned} \tag{14}$$

where c is the context, c_{s1} is the speaker 1 context, c_{s2} is the speaker 2 context, x is the response, z is the latent variable, v_{s1} is the Speaker 1 Cluster Vector, and v_{s2} is the Speaker 2 Cluster Vector. In addition, KL is the KL divergence between distributions and the prior probability distribution, and p, r, and r' are defined as follows:

$$p(z|c,v_{\mathrm{s1}},v_{\mathrm{s2}}) \sim \mathscr{N}(\mu_p,\sigma_p^2), \tag{15}$$
$$r(z|c_{\mathrm{s1}},v_{\mathrm{s1}}) \sim \mathscr{N}(\mu_r,\sigma_r^2), \tag{16}$$
$$r'(z|c_{\mathrm{s2}},v_{\mathrm{s2}}) \sim \mathscr{N}(\mu_{r'},\sigma_{r'}^2), \tag{17}$$

where

$$\left[\mu_p,\log(\sigma_p^2)\right] = \mathrm{MLP}_p(c,v_{\mathrm{s1}},v_{\mathrm{s2}}), \tag{18}$$
$$\left[\mu_r,\log(\sigma_r^2)\right] = \mathrm{MLP}_r(c_{\mathrm{s1}},v_{\mathrm{s1}}), \tag{19}$$
$$\left[\mu_{r'},\log(\sigma_{r'}^2)\right] = \mathrm{MLP}_{r'}(c_{\mathrm{s2}},v_{\mathrm{s2}}), \tag{20}$$

and posterior probability distribution, and q, s, and s' are defined as follows:

$$q(z|x,c,v_{\mathrm{s1}},v_{\mathrm{s2}}) \sim \mathscr{N}(\mu_q,\sigma_q^2), \tag{21}$$
$$s(z|x,c_{\mathrm{s1}},c,v_{\mathrm{s1}},v_{\mathrm{s2}}) \sim \mathscr{N}(\mu_s,\sigma_s^2), \tag{22}$$
$$s'(z|x,c_{\mathrm{s2}},c,v_{\mathrm{s1}},v_{\mathrm{s2}}) \sim \mathscr{N}(\mu_{s'},\sigma_{s'}^2), \tag{23}$$

where

$$\left[\mu_q,\log(\sigma_q^2)\right] = \mathrm{MLP}_q(x,c,v_{\mathrm{s1}},v_{\mathrm{s2}}), \tag{24}$$
$$\left[\mu_s,\log(\sigma_s^2)\right] = \mathrm{MLP}_s(x,c_{\mathrm{s1}},c,v_{\mathrm{s1}},v_{\mathrm{s2}}), \tag{25}$$
$$\left[\mu_{s'},\log(\sigma_{s'}^2)\right] = \mathrm{MLP}_{s'}(x,c_{\mathrm{s2}},c,v_{\mathrm{s1}},v_{\mathrm{s2}}). \tag{26}$$

Here, KL annealing and BoW loss are used for the KL vanishing problem, in which the decoder no longer considers information about the latent variable z as learning progresses. KL annealing weights the value of the KL divergence in Equation 14 linearly from 0 to 1 as the training progresses, and the BoW loss adds a subtask to estimate the set of words in the response from the latent variables to strengthen the association between the latent variables and the words in the response.

Table 2: Automatic evaluation results

Model	Diversity			Similality
	Dist-1	Dist-2	Dist-3	BERT
GVT	0.484	0.720	0.739	0.654
Extended GVT	0.530	0.810	0.821	0.655
Extended GVTSC	**0.640**	**0.950**	**0.975**	**0.672**
Actual response	0.647	0.947	0.963	-

Table 3: Example of generated dialogue responses

Context
Utterance 1: 12.
Utterance 2: What about the 4 times table? (4 段の時は?)
Utterance 3: 16.
Utterance 4: And the 5 times table? (5 段の時は?)
Utterance 5: 24。24,24,24,28. Eh? (24。24,24,24,28。えー。)
Response
GVT: That is right. Yes, so you get it so far. They are the same. (そうそうそう。はい, ここまでいいかな?同じです。)
Extended GVT: Does this formula seem right or wrong? (この式, 合ってそう?違ってそう?)
Extended GVTSC: Does this formula seem right or wrong? (この式, 合ってそう?違ってそう?)
Reference: Is this right? (これで合ってる?)

5 Dialogue Model Evaluation

The acquired elementary school class dialogue data were used as the experimental dataset. The data were partitioned into sub-words using SentencePiece. In terms of the length of the context, the dialogue response was evaluated for up to three turns, and Dist-N [25] and the BERT score [26] were used as automatic evaluation indices. Dist-N is calculated as the ratio of the number of N-gram types to the total number of N-grams, where a higher ratio indicates more diversity. The BERT score uses a pretrained BERT embedding to evaluate the similarity of the response generated by the model and the reference response. The evaluation results for the responses generated by the compared models are shown in Table 2. As can be seen, the extended GVTSC, for all N-grams of diversity outperformed the conventional GVT model. In addition, this was found to be similar to the diversity of actual responses. The similarity evaluation of the extended GVTSC increased by approximately 0.018 compared to that of the GVT model. Here, the consideration of the speaker ' s characteristics during the encoding phase of the encoder is assumed to affect the output vector of each encoder token, which is used to sample the latent variables and attention in the decoder.

Examples of the responses generated when evaluating the elementary school classroom

dialogue data are shown in Table 3. Here, the dialogue was taken from a situation where the teacher asked the question: ”Find the length of the perimeter of a staircase made of squares with one centimeter on each side when it is increased by one step at a time.” As shown in Table 3, in contrast to the conventional GVT model, the proposed extended GVTSC model can generate responses that are related to the context, and these responses are semantically similar to the reference response. In addition, the responses exhibit diversity.

Table 4: Example of extracting similar utterances for class 2 using labels from class 1 data as supervised data

Labeled utterance	Label	Similar utterance	Cos	JW
Thank you. Yes, I will write it here. If you multiply the number on this line by four, it becomes the length of the outside. By the way, are you looking vertically or horizontally?	Mathematical perspective, Functions, Deep learning	○○-san thought about it like this. Focusing on 1 and 15, he/she thought about it 4 times. That is correct. By the way, ○○-san focused on this, but are there people who focused on something else? Yes, you have taken this challenge on.	**0.953**	0.450
段の数を，ありがとう。はい，書くよ。段の数を４倍すると，周りの長さになる。ちなみにさあ，今のは縦に見とるん?横に見とるん？	数学的な見方，関数，深い学び	○○さんはこの考え方でやってやったんだよね。１と１５に目をつけて４倍４倍って考えたんですね。正しいですね。ちなみに○○さんさ，ここに目をつけたんだけど他のところに目をつけた人いる?うん。はいチャレンジャー。		
Yes. With [SEP] what kind of rule did you find? Can you all tell me the rules you found? ○○-kun.	Mathematical perspective, Functions, Deep learning	To find numbers that have not been included in the table, it is better to multiply the numbers with a good cut-off area like 10 [SEP] It does not matter where, but it should have a good cut-off point. OK, well we only have two minutes left.	**0.962**	0.111
うん。[SEP] じゃあ，どんな決まり見つけたん?みんな。。見つけた決まりを教えてください。○○君。	数学的な見方，関数，深い学び	表には表には入っていない値を求めるには，１０のような切りの良い数字で数字から倍すればいいと思いました。[SEP] どれでもいいんだけど切りの良いこの時に，よし。じゃあと２分しかなくなっちゃいました。		

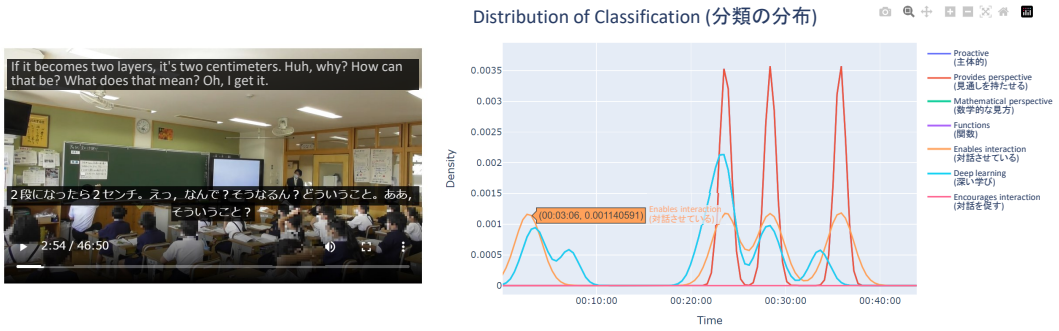

Figure 4: Visualization system for in-class utterance analysis

6 Analysis of Classroom Utterances Using Dialogue Models

6.1 Creating Utterance Vectors using Extended GVTSC

The purpose of this study is to vectorize the dialogue rather than generate dialogue responses using the GVTSC model. First, we used dialogue data to train the extended GVTSC model. We then used the trained model to perform dialogue vectorization. By training the extended GVTSC model on the dialogue, the model can vectorize the utterances in the context required to generate dialogue responses. Finally, dialogue vectors are created by inputting the dialogue as context to the extended GVTSC model and calculating the sum of the CLS token vectors output by the model ' s context TRS encoder and speaker TRS encoder.

6.2 Utterance Analysis using Utterance Vectors

In analyses using the utterance vectors generated by the extended GVTSC model, the classroom dialogue data were created per utterance and as data for multiple utterances dealing with the dialogue. In addition, here, the distance between the utterance and the annotated utterance was obtained. As this distance can be expressed as a vector, the cosine similarity (Cos) was used in this evaluation. For comparison, the Jaro-Winkler distance (JW), i.e., the distance for character matching, was also obtained.

Table 4 shows examples of the extracted similar utterances from the data for class 2 obtained using the labels from the data of class 1 as the supervised data. The first line shows the similarity in single utterance units, from which we can extract similar utterances matching the ”mathematical perspective” (proportional relationship) and ”deep learning” (ask children for other opinions) labels. In addition, the manually added labels for similar utterances are ”mathematical perspective” and ”deep learning” with the results matching the classification by humans. The second line shows the case of two utterance units and extracts the utterance, where the student presents how they found the proportional value. Here, the labeled utterance and surface similarity to the similar utterance was 0.111, which is a very low utterance similarity value. There are no labels that were added manually to this similar utterance, which demonstrates that utterance candidates for applying labels can be obtained mechanically.

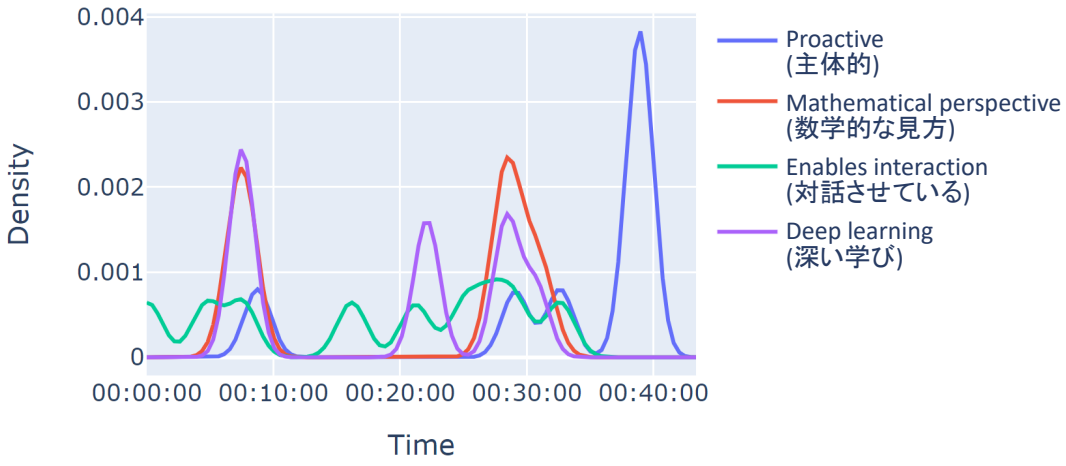

Figure 5: Label estimation and visualization (class 2)

7 Visualization System to Analyze Classroom Utterances

Teachers can receive better feedback by observing the types and tendencies of in-class utterances. Feedback can be made more understandable by visualizing the video of the class, as well as the classification and distribution of the utterances. In this study, we developed a system (Figure 4) that displays the video of the class with utterance subtitles and the distribution of utterance classifications. The distribution of classifications is shown by a graph that uses the probability density function [27][28][29] estimated for each classification.

7.1 Visualizing Learning-related Labels

An example of label visualization is shown in Figure 5. This graph uses the annotated labels for class 2 and estimates and visualizes the class 2 labels. Here, the threshold was set to 0.95 in six utterance units at the time of estimating. Although the annotated utterances are not included in the estimate, these have the same distribution as the manually annotated labels. One difference with the manual labels is that the label ”enabling interaction” is estimated for 20–30 minutes, whereas in the actual class, it matches the part uttered by the children during the presentation.

7.2 Questionnaire-based Evaluation of Visualization System

With open-ended responses, a questionnaire-based evaluation was given to experienced elementary school teachers in relation to a system to visualize learning labels in terms of a time-series distribution. The evaluations by experienced elementary school teachers are shown below.

- A distribution of “deep learning” and “interactive” labels is seen throughout the class as a whole, and this matches classes focusing on “proactive, interactive, and deep learning.”
- If it is possible to check the content of the utterance at a certain point in the distribution, then it is easier to use it as a visualization system.

These are displayed as subtitles in the video; however, to compare with the previous and subsequent utterances, a function can be implemented as a point of improvement in the questionnaire to confirm the content of the utterance. This would enable us to analyze elements based on their relationship to the distribution labels and utterances.

8 Conclusion

In this study, we used a dialogue model for elementary school classes to estimate and visualize learning classifications for utterances in learning. We developed a system to visualize classroom analysis results as a reflection aid to learn classroom trends as a whole through the evaluation of experienced elementary school teachers.

Teachers can employ the visualization system to observe the types and trends of classroom utterances without time-consuming manual reflection, thereby allowing them to work on improving their classes. In addition, teachers can check the classified utterances from the viewpoint of "interactive, independent, and deep learning." Thus, we expect that this will allow us to identify utterances that have a positive influence on students.

In addition, in an extended GVTSC model, in which the proposed utterance information is incorporated, improvements are observed in terms of the diversity and similarity of the generated dialogue responses.

In the future, we can address issues raised in relation to the visualization system obtained from the questionnaire results and the estimation accuracy of the learning classifications.

References

[1] Central Educational Council, "Measures for the comprehensive improvement of teachers' qualification and competence throughout their teaching careers (report)," 2012, (in Japanese). [Online]. Available: https://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/1325092.htm

[2] Ministry of Education, Culture, Sports, Science and Technology, "Courses of study for elementary schools (notification, 2016)," https://www.mext.go.jp/content/1413522_001.pdf, 2016, accessed Oct. 3, 2022 (in Japanese).

[3] K. Akita, *Transforming Pedagogy*. Seorishobo, 2009, ch. The Turn from Teacher Education to Research on Teachers' Learning Processes: Transformation into Research on Micro-Educational Practices, pp. 45–75, (in Japanese).

[4] A. Sakamoto, "A study on teacher reflection processes in lesson study: Focusing on differences between lesson teacher and observing teacher thought processes," *Japan Bulletin of Educators for Human Development*, vol. 9, no. 8, pp. 27–37, 2010, (in Japanese).

[5] T. Yasumori, "Speech protocol analysis during classroom sessions and reflection of elementary school math teachers based on the pck model," *The Bulletin of Japanese Curriculum Research and Development*, vol. 41, no. 1, pp. 59–71, 2018, (in Japanese).

[6] y. Wang, S. Ooi, K. Matsumura, and H. Noma, "Research on a classroom reflection system for improving new teachers' teaching skills," in *Interaction, Information Processing Society of Japan*, 2021, pp. 753–757, (in Japanese).

[7] National Institute for Educational Policy Research, *International Comparison of Teacher Environments: the OECD Teaching and Learning International Survey (TALIS) 2018 Report*, National Institute for Educational Policy Research, Ed. GYOSEI, 2018, (in Japanese).

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[10] O. Vinyals and Q. Le, "A neural conversational model," in *ICML Deep Learning Workshop 2015*, 2015.

[11] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, pp. 3776–3783.

[12] I. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau *et al.*, "A hierarchical latent variable encoder-decoder model for generating dialogues," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/10983

[13] Z. Lin, G. I. Winata, P. Xu, Z. Liu, and P. Fung, "Variational transformers for diverse response generation," *arXiv preprint arXiv:2003.12738*, 2020.

[14] S. Onishi, T. Yasumori, and H. Shiina, "Classroom utterance analysis using a generative deep neural networks for dialogue model," in *2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. Los Alamitos, CA, USA: IEEE Computer Society, jul 2023, pp. 560–565. [Online]. Available: \url{https://doi.ieeecomputersociety.org/10.1109/IIAI-AAI59060.2023.00112}

[15] T. Yasumori, *Examination of teachers' utterances to realize "proactive and interactive deep learning"*, ser. B, Humanities and Social Sciences. Bulletin of Okayama University of Science, 2021, no. 57, pp. 45–52, (in Japanese).

[16] T. Zhao, R. Zhao, and M. Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 654–664. [Online]. Available: https://aclanthology.org/P17-1061

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[19] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[20] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz *et al.*, "Generating sentences from a continuous space," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 10–21. [Online]. Available: https://aclanthology.org/K16-1002

[21] X. Zhou and W. Y. Wang, "MojiTalk: Generating emotional responses at scale," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1128–1137. [Online]. Available: https://aclanthology.org/P18-1104

[22] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji *et al.*, "A neural network approach to context-sensitive generation of conversational responses," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, May–Jun. 2015, pp. 196–205. [Online]. Available: https://aclanthology.org/N15-1020

[23] R. Csáky, P. Purgai, and G. Recski, "Improving neural conversational models with entropy-based data filtering," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5650–5669. [Online]. Available: https://aclanthology.org/P19-1567

[24] B. Sun, S. Feng, Y. Li, J. Liu, and K. Li, "Generating relevant and coherent dialogue responses using self-separated conditional variational AutoEncoders," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5624–5637. [Online]. Available: https://aclanthology.org/2021.acl-long.437

[25] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 110–119. [Online]. Available: https://aclanthology.org/N16-1014

[26] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *In International Conference on Learning Representation*, 2019.

[27] D. W. Scott, *Multivariate density estimation : theory, practice, and visualization*, ser. Wiley series in probability and mathematical statistics. New York ;: Wiley, 1992.

[28] B. W. Silverman, *Density estimation for statistics and data analysis*, ser. Chapman & Hall/CRC monographs on statistics and applied probability. London: Chapman and Hall, 1986. [Online]. Available: https://cds.cern.ch/record/1070306

[29] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.