IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.004, LIIR181
DOI: https://doi.org/10.52731/liir.v004.181

Visualization of the Impact of Classroom Utterances Using Generative Dialogue Models

Sakuei Onishi *, Tomohiko Yasumori †, Hiromitsu Shiina ‡

Abstract

As teachers in elementary school classes have limited time for reflection, it is desirable for the reflection process to automated. Therefore, in this study, we analyze the utterances of teachers and children using a neural network based dialogue model. Additionally, we also analyze and visualize the degree of impact of the utterance during the utterance generation process.

Keywords: Dialogue Analysis, Dialogue Model, Impact of Speech, Speech visualization

1 Introduction

The 2017 Elementary School Guidelines from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) stated that classes should aim to achieve "proactive, interactive, and in-depth learning." Awider variety of initiatives have been taken in elementary school classrooms, and the utterances of teachers and children have also been analyzed in empirical studies in pedagogy. Among these, studies on reflection activities [1] are attracting considerable attention. From proficiency perspective, it is said that teach-ers must be adaptively proficient in developing their students into independent learners [2]. Pedagogical content knowledge (PCK) based reflection studies [3] showed that adaptively proficient teachers demonstrated reflection on the basis of two aspects: instructor-centered and learner–centered PCK, both of which are essential. Currently, in elementary school classes, teachers hold their classes while observing the student's situation, and it is unusual for teachers to teach their classes in a unilateral manner. There are many opportunities for the children to discuss their learning, express their opinions, and provide impressions on the class with their classmates. Thus, student interaction is considered to be a part of the class progression. As a certain degree of interaction is established between teacher and stu-dents, and between the students themselves, it is assumed that if utterances and interactions in the class could be automatically analyzed, then, this would offer significant feedback to teachers. Particularly important elements are classification of teachers' utterances, analy-

* 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

sis of similar utterances, examples of better utterances, and visualization of the impact of teachers' utterances on how children speak during the classes.

We have proposed an extended GVT model [4] for alternating dialogue, among the neural network-based dialogue models. In this study, we propose to analyze classroom dialogue in an elementary school. The dialogue model utilizes an extended Global Variational Transformer Speaker Clustering (GVTSC) model. This model incorporates a clustering mechanism to automatically classify the dialogue in advance, and abstract the speaker's features. The classroom dialogues were recorded from elementary school mathematics classes, and we analyzed the dialogue-style text information transcribed from the recordings. In this analysis, it was possible to use the extended GVTSC to automatically generate utterances after extracting utterances that are in close proximity to a particular utterance. Furthermore, we calculated and visualized the impact of the utterance in the class. The impact of utterances during the class is identified and visualized using two methods. First, by expanding visualization of dialogue relationships between utterances using attention weight, which is a type of Explainable Artificial Intelligence (XAI) method [5], and second, using an improved version of the ERASER benchmark [6] method for assessing validity as a basis for prediction used in dialogue analysis.

2 Classroom Dialogue Data for Experiment

The dialogue data of the classes were obtained by recording a 45-minute math class at an elementary school and transcribing the teacher's and children's speech to create textual information in dialogue form. The data of the two classes are obtained from the 4th grade arithmetic (proportional) class 1 and the 6th grade arithmetic (proportional) class 2. The number of utterances in class 1 is 193, and the number of utterances in class 2 is 274.

3 Extended GVTSC Model with Added Speaker Clustering

3.1 Overview

When generating dialogue responses, as safe responses can be generated as responses to various dialogues, the diversity of responses may decrease [7]. The Global Variational Transformer (GVT) model [8] uses sampled latent variables as input to a decoder. It is assumed that diversity of response can be achieved by expressing and sampling the characteristics of the speaker with latent variables. However, in previous research, it has been shown that this tends to decrease the consistency of response generated by latent variables [9]. Therefore, to consider the characteristics of each speaker, the characteristics of each speaker are abstracted using clustering and an encoder takes into account the characteristics of the speaker to improve both consistency and diversity. In this study, an extended GVTSC model is proposed, in which clustering that classifies the speakers in advance is added and evaluated.

3.2 Creating Speaker Characteristics through Clustering

An overview of the extended GVTSC model is shown in Figure 1. In the extended GVT model, an encoder was added to the GVT model for each speaker. To this, we added a part that uses clustering to create a feature vector for the speaker (the dotted line in Figure 1), and this is used in context encoding. Here, we shall explain the processing of the extended

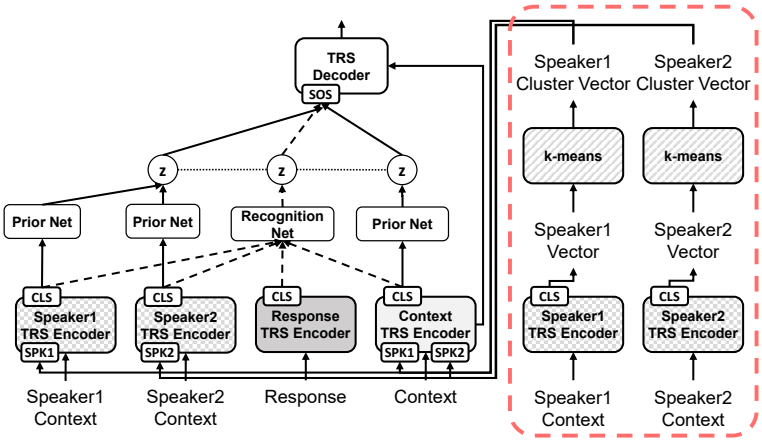

Figure 1: Structure of extended GVTSC

GVTSC model. Firstly, clustering is used to create the feature vector for the speaker. The context is a summary of the utterances of the two interacting parties, and this can be divided for each speaker. Therefore, the context of the dialogue is divided between speakers, and processing occurs for each speaker. As the processing is the same for each speaker, let us describe the situation in the case of Speaker 1. The Speaker 1 context is encoded using Speaker 1 TRS Encoder. With the 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 Speaker 1 (Speaker 1 Vector). Clustering is performed in relation to Speaker 1 Vector. In this study, k-means is used for clustering. The number of clusters k and the hyperparameters needs to be experimentally determined. As a result of clustering, the clusters belonging to the Speaker 1 Vector are predicted and the central vector for this cluster (Speaker 1 Cluster Vector) is obtained. The same processing is performed for Speaker 2 as for Speaker 1, and the Speaker 2 Cluster Vector is 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.

3.3 Dialogue Response Generation

The whole context of the dialogue is input into 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; Speaker 1 Cluster Vector is input into SPK1, and Speaker 2 Cluster Vector is input into SPK2. Furthermore, the context is divided for each speaker, and each is input into the respective Speaker TRS Encoder to obtain the output vector. Meanwhile, the Speaker 1 token (SPK1) is added to the input sequence in Speaker 1 TRS Encoder and the Speaker 1 Cluster Vector while 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 uses the feature vector of each speaker, the encoding aims to consider the characteristics of the speaker. The sampling of latent variable z from Prior Net and Recognition Net is approximated by multilayer perceptrons (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, Prior Net estimates the mean and variance of the context vector using MLP. The latent variable z is sampled from the normal distribution following this mean and variance. With Recognition Net, in addition to the Speaker TRS

Table 1: Automatic evaluation results

Model	Diversity			Simirality
	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	-

Encoder and Context TRS Encoder, the output vector of the Response TRS Encoder CLS token is used to estimates the mean and variance of the entire dialogue vector using 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 seen as 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 TRS Decoder, the latent variable of the response speaker is inputted in addition to the normal latent variable to the SOS token at the beginning of the input sequence, and the latent variable is used for generating the response.

4 Evaluating the Dialogue Model

Elementary school class dialogue data was used for the dataset. As preprocessing, this was partitioned into subwords using SentencePiece. In terms of the length of the context, the dialogue response was evaluated for up to three turns. Dist-N (Li et al., 2016) and BERT Score [10] 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 and is an indicator wherein the higher the ratio, the greater the amount of diversity. The BERT Score is a method that uses pre-trained BERT embedding to evaluate the similarity of the response generated by the model and the reference response.

The results of the response generated by each model using the automatic evaluation indices are shown in Table 1. In the clustering (k-means) of the extended GVTSC model, the number of clusters is set to 8 based on the results of preliminary experiments. The extended GVTSC, for all N-grams of diversity, is evaluated as being higher than that of the GVT model. Furthermore, this is found to be close to the diversity of reference responses. The similarity evaluation via extended GVTSC is found to be increased by approximately 0.018 compared to that of the GVT model. Consideration of the characteristics of the speaker during the encoding phase of the encoder is assumed to affect the output vector of each encoder token, which is used for sampling latent variables and attention in the decoder.

Examples of the generated responses when evaluating the elementary school classroom dialogue data are shown in Table 2. Table 2 shows the dialogue in 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”. In contrast to the GVT model, the extended GVTSC model can generate responses related to the context, and these responses are semantically similar to the reference response. Additionally, these are diverse responses.

Table 2: Example of dialogue response generation

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 Creating Utterance Vectors using Extended GVTSC

The purpose of this study is not to generate dialogue responses using the GVTSC model but to vectorize the dialogue. First, we used dialogue data to train the extended GVTSC model. Next, we used the trained model to perform dialogue vectorization. By using the dialogue data and training the extended GVTSC model, the model achieved the ability to vectorize utterances in the context required to generate dialogue responses. Finally, dialogue vectors are created by inputting dialogue as context to the extended GVTSC model and calculating the sum of the CLS token vectors output by the extended GVTSC model Context TRS Encoder and two Speaker TRS Encoder.

6 Utterance Analysis using Utterance Vectors

In analyses using generated utterance vectors by the extended GVTSC model, the classroom dialogue data is created per utterance and as data for multiple utterances dealing with the dialogue. Furthermore, here, the distance between the utterances is obtained. As this distance can be expressed as a vector, Cosine similarity (Cos) is used. For comparison, the Jaro–Winkler distance (JW), which is the distance for character matching, is also obtained. Example of extracting similar utterances from the Class 2 data for the Class 1 data are shown in Table 3. In the first line, the similarity is measured in units of one utterance, and similar utterances that match the ”proportionality relation” and the ”asking other children for their opinions” are extracted. The second line is the case of two utterances, and extracts the utterance in which the child announces how to find the proportional value. The superficial similarity between the utterances is 0.111, which is very low, but the Cos similarity using the extended GVTSC model is as high as 0.962, suggesting that the vector similarity using the dialogue model is effective.

Table 3: Example of Extracting Similar Utterances from Class 2 for Class 1

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。じゃああと２分しかなくなっちゃいました。		

7 Analysis of the Impact of Utterances on Dialogue Response

In the field of XAI, research to provide evidence for the output of AI is being conducted. Regarding the XAI techniques for natural language processing, there are methods that visualize the attention weight in models using attention, such as Transformer, and use them as grounds for prediction. Also, methods that infer the important input parts by quantitatively analyzing in what way the output changes when changing the input have been proposed. One such method is the ERASER benchmark that evaluates reasonableness as grounds for prediction. Herein, we analyze the impact of classroom utterances by modifying both methods for the extended GVTSC model proposed in this study.

7.1 Visualization Method using Attention Weight

In this study, the visualization of attention weight is possible using the Transformer within the extended GVTSC model. While generating dialogue responses, weights for each token in the context are computed internally within the model to generate a given dialogue response. Hence, it is possible to analyze the dependency relationships of important context tokens by using attention weight. Therefore, by visualizing attention weight in the dialogue model, analysis of important token, as well as utterance analysis, using token and utterance units are performed. The method of visualizing attention weight consists of the following steps:
(1) The weight on each response token is averaged, and the weight on each context token is calculated,
(2) The weight on each context token is normalized to a value between 0 and 1,
(3) The background color of the token is highlighted in red in accordance with the weight magnitude.

In particular, the attention weight obtained from the model is the weight between each response token and context token, and it does not measure the impact of the response on the entire utterance. Therefore this is converted into the weight on the response utterance as a whole during step (1).

In the analysis of response utterances, the neural network dialog model is forced to generate the utterance response utterances to be analyzed.

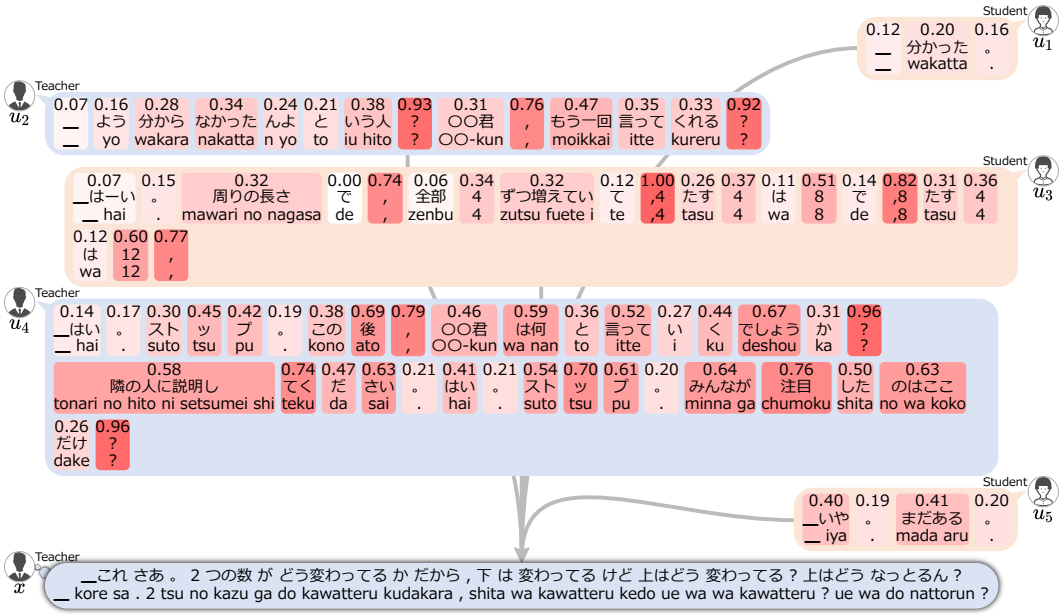

Figure 2: Example of Visualization of Attention Weights (per Token)

Figure 2 shows how attention weight changes in relation to the utterance of the teacher - “Here’s how the two numbers are changing, so the bottom’s changing, but how’s the top changing? What’s going on up there?” is visualized. The speaker (teacher, child) is assigned at the beginning of the utterance, and attention weights and tokens are paired at the top and the bottom. The final utterance is a response utterance generated by the dialogue

model, and the impact is represented by arrow lines going from the context to the response utterance. The important tokens in relation to the response utterance can be interpreted based on the attention weight. In the example in Figure 2, “?,” numbers, “attention,” etc., have a high weight, and therefore this can be interpreted in relation to the response utterance concerning changes in the numbers. An example of visualization of attention weights per utterance for the same dialogue as in the per-token example is shown in Figure 3. The weight of the fourth utterance is the highest at 1.00 in Figure 3, and this suggests the influence of the fact that the overall weight of the utterance is high in addition to “?” and “attention.” Additionally, the third utterance related to the changes in the numbers is the third biggest number, while the first utterance is the smallest number, and this is inferred to be impacted by the fact that the information volume is low.

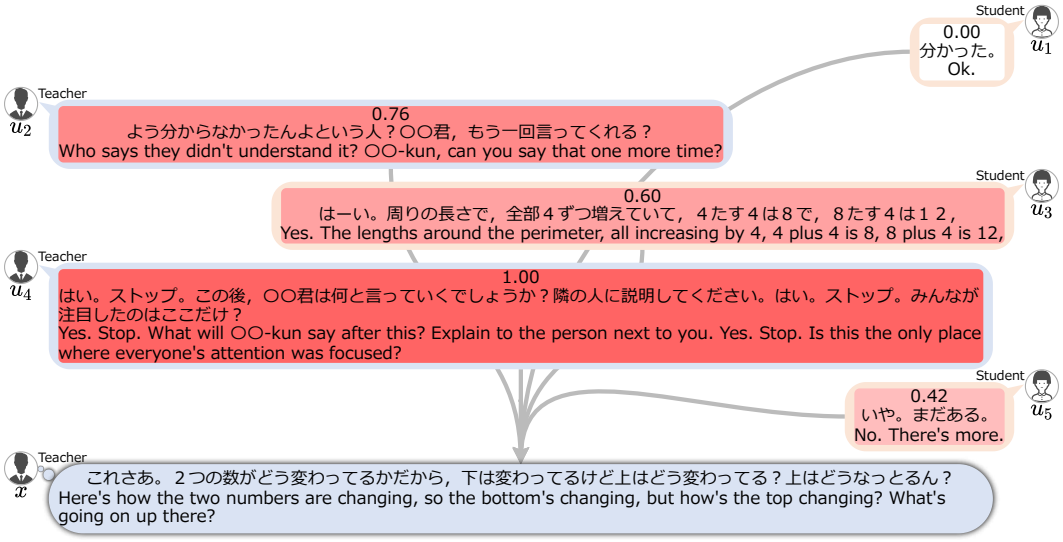

Figure 3: Example of Visualization of Attention Weights (per Utterance)

7.2 Analyzing the Impact of Utterances using the Generation Probability of Dialogue Responses

ERASER benchmark has been proposed as the benchmark to assess the reasonableness of using them as evidence. ERASER proposes datasets used in the evaluation as well as the hard and the soft types of evaluation methods.

Furthermore, with the soft type, it is possible to evaluate continuous scores such as attention weights, and the two types of evaluation indicators of comprehensiveness and sufficiency have been proposed. In this study, an improved Dialogue Dependency index was used to analyze the impact of comprehensiveness on utterances in dialogues.

Comprehensiveness in ERASER is an index that states that lower the predictive probability, higher is the comprehensiveness of the evidence, in case the evidence is excluded from the input. In this study, the Dialogue Dependency is used to remove utterance from dialogue context, wherein the lower the predictive probability of the removed utterances, the more important it is. The probability $p(x|c \setminus u_i)$ that a response x is generated from a dialogue context $c = u_1, u_2, \ldots, u_N$ excluding utterance u_i and the probability $p(x|c)$ that a

response x is generated from the original dialogue context c is used to define the Dialogue Dependency index DD_i for the utterance u_i as follows,

$$DD_i = p(x|c) - p(x|c \setminus u_i).$$

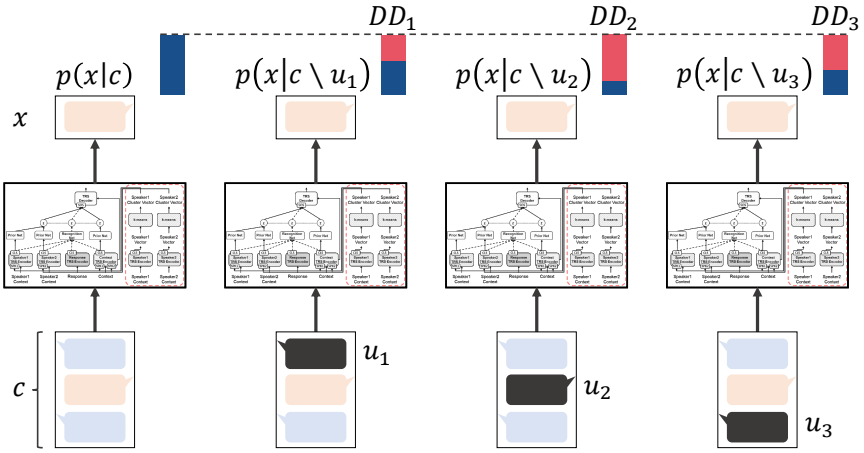

Figure 4: Method for Analyzing the Impact of Utterances using the Generation Probabilities of Dialogue Responses

Here, we shall provide an overview of the method for analyzing the impact of utterances using dialogue response generation probability. The flow of incorporating the dialogue context c into the model and generating the model response x is shown in Figure 4. The first item from the left shows the probability $p(x|c)$ that a response x will be generated from the context c of the original dialogue. Items $2-4$ are the probabilities $p(x|c \setminus u_i)$ of the utterances u_i being removed from the dialogue context c and the response x generated. The DD_i, which determines the degree of impact the utterance u_i, is obtained from the difference between the original generation probability $p(x|c)$ and the generation probability $p(x|c \setminus u_i)$ that excluded the utterance u_i.

Figure 5 shows an example of visualizing the Dialogue Dependency indices for an utterance using the generation probability. The left side shows the dialogue between the teacher and the students. On the right side, we can see the generation probability $p(x|c \setminus u_i)$ as a bar graph and the DD_i. The top section on the right is the probability $p(x|c)$ of a response x being generated from the original dialogue context c that we can see below it the probability $p(x|c \setminus u_i)$ of the response x being generated from the dialogue context c excluding the utterance u_i and the DD_i. Here the removed utterance u_i corresponds to the utterance on the left side of the bar graph. In Figure 5, we have visualized the Dialogue Dependency indices for utterances using the generation probability in relation to the same dialogues for examples in Figure 2 and Figure 3, in which the attention weight is visualized. For the response utterance regarding the change in number, we can see that the DD_3 for the third utterance u_3 in which the change in the number is discussed is high at 0.477, and when this utterance is excluded, the generation of the response is extremely problematic. In contrast to this, the first, second, and fifth utterances have low Dialogue Dependency indices, and the impact on the response utterances generation is low. Particularly, in the fifth utterance u_5, the Dialogue Dependency index DD_5 is a negative value, and excluding this utterance

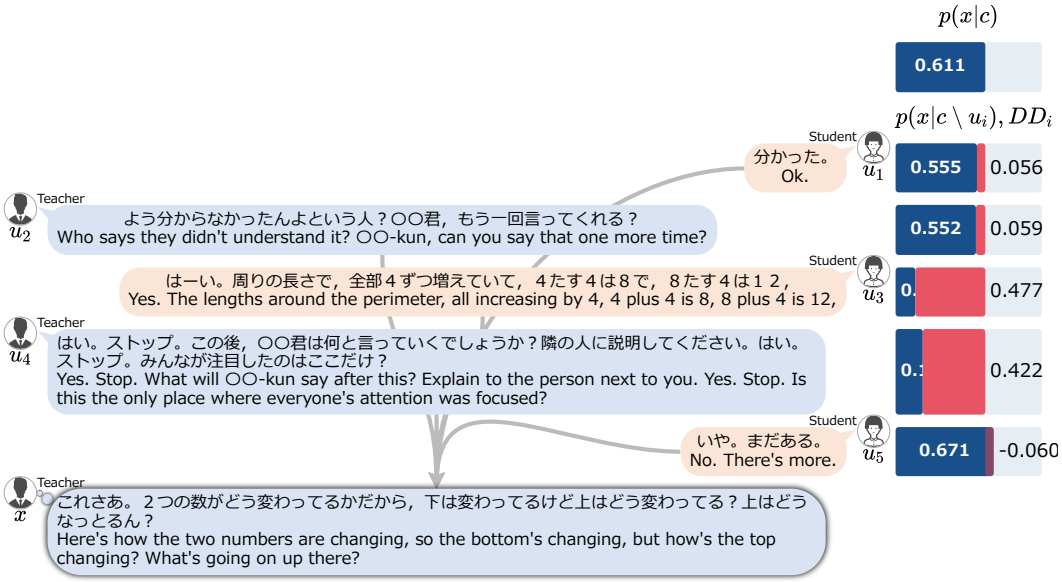

Figure 5: Example of Visualization of Dialogue Dependency Indices of Utterances using Generation Probabilities

leads to high probability of generating a response utterance, indicating that this utterance may just be noise.

7.3 Comparison of Analysis Methods

Here, we can compare the method using attention weight and the method using dialogue response generation probability. First, from the perspective of calculation cost, the method using attention weight is able to obtain attention weight by generating a response from the dialogue context. In contrast, as the method using dialogue response generation probability generates responses by excluding utterances from the context of the dialogue in order, the number of generations required is the number of utterances in the context in addition to the generation of responses from the dialogue context. In terms of computational cost, the method using attention weight is superior. Next, a comparison is made in terms of accuracy when determining the impact of utterances. If we compare the evaluations of the examples of both methods, the method using attention weight only has a low evaluation for the first utterance. On the other hand, the method using the dialogue response generation probability has a low evaluation for each of the first, second, and fifth utterances. Thus, there is a difference between the results of these methods. In terms of relevance to the response utterances regarding the change in number, the first, second, and fifth utterances are low, suggesting that the method using the generation probability of the dialogue response is more accurate.

8 Conclusion

In this study, we analyzed classroom utterances using a dialogue model in an elementary school class. In particular, we analyzed how the teacher's and students' utterances in the elementary school class impacted the later utterances, using the attention weight visual-

ization method and a visualization method based on the generation probability of dialogue response. The proposed extended GVTSC model incorporates speaker information, and demonstrated improvements in diversity and similarity in generating a dialogue response. Moving forward, future challenges include improving performance of the dialogue model used for utterance analysis, collecting class data, and improving the utterance analysis method.

References

[1] 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).

[2] A. Sakamoto, "How do in-service teachers learn from their teaching experiences?" *The Japanese Journal of Educational Psychology*, vol. 55, no. 4, pp. 584–596, 2007, (in Japanese).

[3] 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).

[4] T. Onishi, S. Onishi, and H. Shiina, "Improved response generation consistency in multiturn dialog," in *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2022, pp. 416–419.

[5] A. Madsen, S. Reddy, and S. Chandar, "Post-hoc interpretability for neural nlp: A survey," *ACM Comput. Surv.*, vol. 55, no. 8, dec 2022.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.