# Online Discussion Transition Analysis for Group Learning Support

Yasunobu Sumikawa
*Takushoku University*
Tokyo, Japan
ysumikaw@cs.takushoku-u.ac.jp

Akikazu Takada
*Ddrive K.K.*
Ehime, Japan
takatter@ddrive.ai

Akira Ichinose
*First Torrent LLC.*
Fukuoka, Japan
ichinose@first-torrent.com

Ari Murakami, Yuki Toyono
*Ddrive K.K.*
Ehime, Japan
{ari,yuki}@ddrive.ai

Ryohei Ikejiri, Meiko Sakasegawa
*The University of Tokyo*
Tokyo, Japan
{ikejiri,sakasegawa}@iii.u-tokyo.ac.jp

Kaoru Sekine
*Tohoku University*
Miyagi, Japan
kaoru.sekine@gmail.com

Yuhei Yamauchi
*The University of Tokyo*
Tokyo, Japan
yamauchi@iii.u-tokyo.ac.jp

***Abstract*—Group learning has been recognized as an effective learning method. Recently, group learning using online communication tools and face-to-face group learning have been increasing. This change in the learning environment has highlighted the importance of supporting teachers in monitoring the learning progress of each group. In this study, we propose an algorithm to analyze the discussion transitions. The algorithm first extracts the entities and Wikipedia categories from the discussion text. It then applies a queue model, which limits the analysis to the most recent topic, or a memory model, which emphasizes repeated topics. Subsequently, it triggers another algorithm that allocates a Wikipedia category as the main theme for each discussion text. To evaluate the effectiveness of the proposed algorithms, we tested them on the W2E dataset, which includes topics comprising multiple events and an actual conversational dataset. The test results confirmed that the algorithms analyzed the discussion transitions and main themes with high accuracy.**

***Index Terms*—Discussion analysis, transition analysis, topic detection and tracking, Wikipedia**

## I. Introduction

The educational technology and learning science fields have recognized group work as an effective method for higher-order learning. In conjunction with this research finding, the Organization for Economic Co-operation and Development and Japan's new guidelines [6] urge group work in classrooms.

Recently, opportunities for group learning have been increasing through face-to-face group learning and the use of online communication tools. Unlike face-to-face classes, online tools use mechanisms, such as breakout rooms for group activities, which limit the participation of non-group members. While these mechanisms assist in group activities, teachers find it difficult to monitor all the groups simultaneously. Thus, providing support for group learning and teachers is becoming increasingly important.

This study aimed to analyze whether the discussions in each group are appropriately transitioning and what themes the current discussions are. These analyses will help teachers understand the challenges participants face during group learning. Below are some examples of the kind of discussions in which teachers would intervene in this study. In these examples, we assume two students, A and B, discussing the difference between birds and dinosaurs.

A: Let's find out the differences between birds and dinosaurs.
B: I will look into dinosaurs.
A: Then I will look into birds.
A: There is a difference between living and extinct.
B: Sure. In addition, are there any differences in the skeletons of dinosaurs and birds?
A: Good. Let's look into the differences between their skeletons now.
B: Speaking of which, what's the next class?
A: I don't know.
B: Oh, a recent study reports about the characteristics of dinosaurs.
A: I'm reading those research results, but I don't understand.
B: I wonder what these results are.

The above discussion was congruous with the theme until it focused on skeletal differences. However, after deciding to research independently, B remarked, "what's the next lesson?" This question is irrelevant to the discussion. Then, B found the latest research results, and the discussion returned to the original theme of the conversation. However, both A and B had difficulty understanding the report content, thus impeding further discussion. For such cases, this study proposes consulting a teacher so that the group can continue the discussion on birds vs. dinosaurs in two scenarios: (1) when the discussion moved inappropriately, as in the case of the question, "what's the next lesson?" and (2) when the discussion stalled because neither understood the research result. In this example, we assumed that in the former case, the conversation between the two students returned immediately to the theme. If the situation had persisted, teachers would have been required to address another challenge of avoiding wasting learning time.

The authors would like to acknowledge Google Inc. for supporting this project through the "Research on developing a system for detecting the groups that need group work support."

We assumed that *although the discussion transition happens over time, the change is continuous.* To represent this assumption, we propose a queue model that limits the analysis to the most recent texts, and a memory model that emphasizes repeated topics. In addition, we present two algorithms to analyze the main themes of the discussion texts. The first finds the center of gravity after mapping the Wikipedia categories obtained from texts to the coordinate space. The second considers the important themes repeated over time.

When a discussion text is given, the output of our algorithms are a numerical value indicating that the discussion is appropriately transitioning and a Wikipedia category representing the main theme. These results would help teachers find groups that require intervention.

We applied the proposed algorithms to the W2E dataset to evaluate their effectiveness. This dataset includes topics comprising multiple events and actual conversational data. The test results confirmed that our algorithms correctly analyzed the discussion transitions for 98% of the texts. They allocated appropriate Wikipedia categories as the main themes for discussion texts with 70% accuracy. We created a subset of deviating and stalling discussions from the W2E dataset and found that the proposed algorithms could correctly detect all of them. Finally, we performed the same analysis on actual conversational text to confirm they analyzed of the discussion transition properly.

The remainder of this paper is structured as follows. In the next section we present related work. In Section III, the proposed algorithm. Section IV provides the experimental evaluation results with detail analyses. Finally, the last section concludes the paper and outlines future work.

## II. Related Work

### A. Discussion Analysis

Research has conducted on supporting teachers for group learning. Taoufiq *et al.* proposed a system using latent Dirichlet allocation (LDA) to visualize learners' discussions [8]. The system applies LDA to each discussion text and displays the words that constitute the topic with the highest probability, resizing them by their probabilities in the UI.

Taoufiq *et al.*'s study aimed to assist teachers in providing appropriate feedback to all groups. By contrast, this study aimed to discover groups in which teachers should intervene. As the objectives of these two studies are different, they can be used simultaneously. For example, if no group requires the teacher's intervention the system proposed by Taoufiq *et al.* could assist the group in better discussions. However, if the discussion is not going well, the proposed algorithm can help teachers promptly join the group in the middle of a session, organize the issues, and encourage learners to discuss.

### B. Topic Detection and Tracking

This study analyzed whether each group was adequately engaged in the discussion. This evaluation is similar to analyzing whether an entered text has the same content as the previous texts. Such analysis has been studied in machine learning as topic detection and tracking (TDT). Radinsky and Davidovich proposed a TDT algorithm for predicting future events after mining causal relationships [7]. Their algorithm analyzes whether the entered news is on the same topic as previously reported. If it is the same, the algorithm associates this latest news with the chain it has accumulated as news.

This procedure is identical to the discussion transition analysis in this study. If the entered discussion text is the same as the previous content, our algorithm also accumulates new text. Otherwise, the algorithm discards the newly entered text.

## III. Algorithm

Fig. 1 shows an overview of the proposed algorithm. Before applying the algorithm, we assumed that the learner's discussion had already been converted to text[1]. The algorithm first extracts the entities and Wikipedia categories from the text. These are used in the discussion transition analysis that reveals three types of inputs: appropriately transitioning discussion, deviating discussion, and stalling discussion. If the algorithm determines the input text as appropriately transitioning discussion, the algorithm records the result so that for subsequent analyses. Subsequently, the algorithm returns a numerical value indicating a recommendation for the teacher to intervene. Finally, the algorithm analyzes the main theme of each text to identify the ongoing discussions of each group easily. The two results are then presented on the screen for the teacher to review.

The following sections describe the processes of extracting entities and Wikipedia categories from discussion texts. We then describe the discussion transition and thematic analyses.

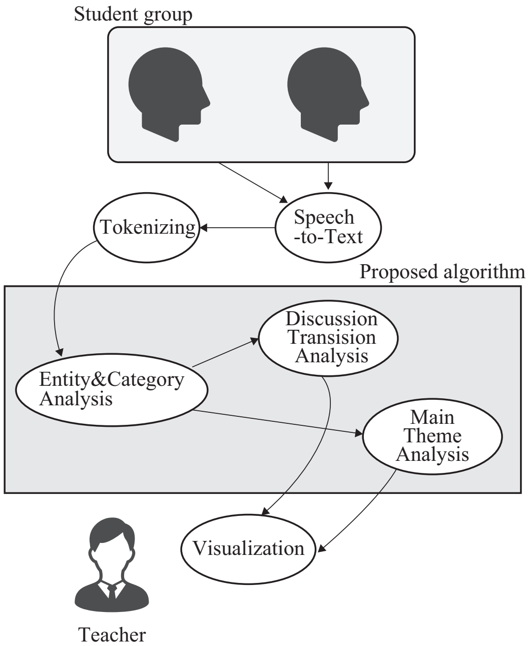

Fig. 1. System overview

[1] Our implementation applies Google Cloud Speech-to-Text to obtain discussion texts. https://cloud.google.com/speech-to-text

### A. Entity & Category Analyses

In this study, we have used two terms: entity and Wikipedia category. The former represents those used in the text, whereas the latter represents the meaning of the text. We can use any existing tools for entity detection[2]. To map descriptions from Wikipedia, we employ explicit semantic analysis (ESA) [2], which outputs Wikipedia articles ranked based on their relevance to the input text. ESA helps retrieve Wikipedia articles highly relevant to the discussion text. Wikipedia articles have several categories defined by their editors. We regarded these categories as abstractions of their articles. In other words, this Wikipedia category set is considered the topic of the input discussion text. However, Wikipedia categories are assigned if they are relevant to the article; some categories are unrelated to the discussion. To remove such noise, we analyzed only those categories containing nouns extracted from the discussion text. The entity and Wikipedia categories produced from this process were aggregated into a word list $words$ for further analysis.

### B. Discussion Transition Analysis

This study proposes two models for analyzing discussion transitions: queue and memory models. Both models assume that themes change over time. However, these two models function differently to capture this change. The queue model excludes old text from the analysis. The memory model assigns weights to new analysis results assuming *topics change over time; however, repeatedly appearing topics are central to the discussion content.* To implement this assumption, our model reduces these values over time. We describe the two models in the following sections.

*1) Queue model:* Queue is a type of data structure that removes objects in the First In First Out order. It analyzes the discussions from the most recent texts by storing these texts and excluding the oldest ones.

This model limits the analysis of the discussion transition to the most recent N pieces. It assumes a uniform distribution, implying the analysis is equal for N texts. In other words, the model analyzes whether the new text matches any N previous results. If the number of common results is below a threshold $\alpha$, the discussion is considered deviating. In contrast, if the common number is greater than a threshold $\beta$, the discussion is considered stalled.

Eq. 1 shows the formal equation of this model.

$$QVal(words, N) = \frac{\sum_{w \in words} \delta(w, Words(N))}{\mid words \cup Words(N) \mid} \tag{1}$$

where $Words(N)$ is a function that retrieves words from N past discussion texts. $\delta(w, Words(N))$ denotes a function that returns 1 if $w$ is included in $Words(N)$; otherwise, it returns 0. The numerator of this equation counts the number of common words between the past and the new texts. As the denominator represents the number of words in all analyzed texts, this division expresses the number as a probability. In the queue model, if the value of this equation is between $\alpha$ and $\beta$, new text is added to the discussion text, and the oldest text is excluded.

*2) Memory model:* This model assumes that content discussed repeatedly for a long time is more important than the most recent content. We represent this assumption by modeling the memory forgetting curve [3], which describes how content is retained through repeated learning.

Fig. 2 shows the weight calculation for words included in $words$. It shows the weight calculation of three texts generated at times t, t', and t" containing the words "soccer", "sports", and "Japan". The model reduces the weight of each word, exponentially over time. However, the weights for each word are summed so that words that appear repeatedly have a higher weight. We use the exponential function $exp$ to calculate the weights according to the forgetting curve of memory model. If the difference between the current time $t$ and the time $t'$ of the previously posted text increases, the weight of the word decreases. Thus, this difference is given as the argument of $exp$. The following equation shows the proposed method for calculating the weights of a single text using:

$$f(t, t') = exp^{(-(t-t'))} \tag{2}$$

Each time a new text is entered, the weights are recalculated because the already parsed text becomes older than before. Additionally, because the same word is assigned different weights in each text, the sum of these weights is the final weight of that word at time $t$. We calculate the sum of Eq. 2 to find words that represent the current themes from the already parsed text, as shown in the following equation:

$$W(w, t) = \sum_{t' \in PText(t)} f(t, t') \times \delta(w, text(t')) \tag{3}$$

where $w$ is a word. The function $PText(t)$ retrieves the time recorded before time $t$. $\delta$ is a function that returns 1 if $w$ is used in the $text(t')$; otherwise, it returns 0.

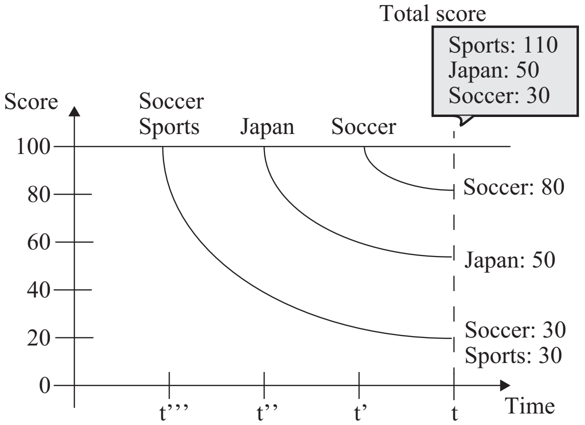

Fig. 2. Memory model

[2] We use AutoML in our implementation. https://cloud.google.com/natural-language/automl/docs

Finally, we sum up the analyzed results of all previous texts. The final result is calculated using the following equation:

$$MVal(words, N) = \sum_{t \in N} \sum_{w \in words} W(w,t) \tag{4}$$

Similar to the queue model, we compared this score with the two thresholds, $\alpha$ and $\beta$, to determine the results. The memory model compares the score of the text in $PText(t)$ with the two thresholds so that it is analyzed at time $t$.

### *C. Main Theme Analysis*

To analyze the main theme of each text, we select the appropriate Wikipedia categories obtained by ESA. Wikipedia categories were originally introduced to organize Wikipedia articles; however, these categories are considered as a summary. Using this feature, we regard the Wikipedia categories obtained for a discussion text as abstractions and select the main theme from them. This selection method uses a different procedure for each discussion transition model to quantify each category and select the appropriate one as the main theme.

*1) Queue model:* The queue model uniformly analyzes the most recent text, thus, embedding all categories into the coordinate space. Each Wikipedia category is assigned to one or more Wikipedia articles. We regard articles belonged to this category as its describing texts. We first train Doc2Vec [5] using all Wikipedia articles to embed these categories. We then apply the Doc2Vec model to texts of all Wikipedia articles, collated as one document from each category. This process allows us to convert each Wikipedia category into a feature vector. We then compute the center of gravity. Subsequently, we regard the Wikipedia category that is the closest to the center of gravity in coordinate space as the main theme.

*2) Memory model:* The memory model computes a score for each Wikipedia category using Eq. 4. Wikipedia categories repeatedly appearing over a long period are assigned a higher score. Thus, we regard the Wikipedia category with the highest value obtained from Eq. 4 as the main theme.

### *D. Algorithm Overview*

Algorithm 1 shows the pseudo code for the proposed algorithm. It first extracts entities and Wikipedia categories from the text input *txt* (lines 2∼4). Then, we apply one of the two models, queue or memory, described in Sec. III-B to integrate these results in line 5 to obtain the score for all discussion texts (lines 6∼7). If this score is between the thresholds, we consider the discussion progressing properly and record it as the discussion text (lines 8∼10). Finally, lines 11∼12 show the analyzed main theme of the input text. This result and the probability values of the discussion transition are returned as the results of proposed algorithm.

**Algorithm 1** Algorithm overview

```
Input: A text $txt$, discussion texts $N$, thresholds $\alpha, \beta$
Output: A score $score$, main theme $m$

1:  Function $DiscussionAnalysis(txt, N, \alpha, \beta)$
2:    // Text Analysis
3:    $entities \leftarrow ExtractEntities(txt)$
4:    $WikiCats \leftarrow ESAbasedWikiCatExtraction(txt)$
5:    $words \leftarrow entities + WikiCats$
6:    // Discussion Transition Analysis
7:    $score \leftarrow$ a result of Eq. 1 or Eq. 4
8:    if $\alpha < score < \beta$
9:      $N \leftarrow txt$ // Store $txt$ to discussion texts
10:   end if
11:   // Main Theme Analysis
12:   $m \leftarrow$ results of Sec. III-C
13:   return $score, m$
```

## IV. Evaluation

### *A. Research Questions*

In this study, we performed experimental evaluations based on the following five research questions:

- **RQ1**: Can our algorithm detect properly transitioning discussions?
- **RQ2**: Can our algorithm detect stalling discussions?
- **RQ3**: Can our algorithm detect deviating discussions?
- **RQ4**: Can our algorithm assign the Wikipedia category as the main theme for text?
- **RQ5**: How well does our algorithm work for actual conversational texts?

### *B. Experimental setting*

*1) Dataset for **RQ1** and **RQ4**:* As **RQ1** and **RQ4** are TDT problems, we performed quantitative evaluations on the W2E dataset [4] created for the TDT problem. W2E dataset includes topics comprising events reported by news agencies, such as Reuters, the New York Times, and BCC. This dataset had topics from news reports of 2016; however, 3,083 topics are available. The events are stored in the English version of Wikipedia. These events were manually aggregated with multiple related events as topics.

The W2E dataset was created as a ground-truth TDT dataset; however, some topics includes only one event. We selected topics containing more than 3 events as we analyze the topic transition of multiple texts. After the filtering, we collected 269 topics containing 1,781 events from the W2E dataset. In our experimental evaluation, each topic had one of the following 9 categories: Sport (**S**), Armed conflicts and attacks (**AA**), Business and economy (**BE**), Arts and culture (**AC**), Law and crime (**LC**), Politics and elections (**PE**), International relations (**IR**), Disasters and accidents (**DA**), and Health and medicine (**HM**).

Table I shows the statistics of the filtered W2E dataset. Table II shows the average number of events, topics, and tokens in the texts for each category. Additionally, we opened the filtered W2E dataset used in this study for re-examination[3].

[3] https://onl.tw/NnFiEXM

TABLE I
STATISTICS OF TEST DATASET.

| | |
|---|---|
| Num. of topics | 269 |
| Num. of events | 1,781 |
| Ave. num. of tokens | 32.9 |
| Ave. num. of events per topic | 6.62 |
| Num. of subjects for **RQ5** | 7 |
| Num. of characters in conversational text for **RQ5** | 7,081 |

TABLE II
STATISTICS OF TEST DATASET BY CATEGORY.

| | **S** | **AA** | **BE** | **AC** | |
|---|---|---|---|---|---|
| Ave. Num. of events | 4.4 | 10.0 | 5.8 | 3.2 | |
| Num. of topics | 13 | 69 | 9 | 4 | |
| Ave. Num. of tokens | 35.5 | 30.3 | 29.0 | 45.4 | |
| | **LC** | **PE** | **IR** | **DA** | **HM** |
| Ave. Num. of events | 5.7 | 4.8 | 4.8 | 9.8 | 8.0 |
| Num. of topics | 24 | 73 | 56 | 15 | 6 |
| Ave. Num. of tokens | 34.6 | 32.6 | 38.4 | 34.3 | 28.4 |

*2) Dataset for **RQ2**:* To evaluate **RQ2**[4], we extended the W2E dataset to simulate stalled discussions by duplicating events in each topic. This evaluation used 269 topics as well as **RQ1**.

*3) Dataset for **RQ3**:* For the evaluation of **RQ3**[5], we extended the W2E dataset again by inserting an event of a topic to the end of another topic to produce unrelated and unexpected stories. We performed this operation for all two topic combinations. However, to ensure that two completely unrelated topics were mixed, mixing was discontinued if one of the following conditions was satisfied:

- If the text of the event to be added and the text of all the events of the topic have common words.
- The feature vector created by the latent semantic analysis (LSA) for the texts of the event to be added depends on the feature vector created by the LSA for the texts of each event of the topic.

To test the second condition, we used the mutual information shown in the following equation:

$$MI(A,B) \quad = \quad \sum_{a \in A} \sum_{b \in B} p(a,b) \log \left( \frac{p(a,b)}{p(a)p(b)} \right) \qquad (5)$$

The higher the measurement score, the more the correlation between the events. This procedure produced 2,055 topics.

*4) Dataset for **RQ5**:* After analyzing the research questions using Wikipedia texts, we evaluated the proposed algorithm for the actual conversational data. We conducted the final analysis with Japanese speakers. They discussed topics from several disciplines, such as science, social studies, the Japanese language, and information science, along with topics from Japanese elementary, junior high, and high schools. Each discussion was conducted for 10 minutes between two people. The conversational data were converted to text using Google Cloud Platform's speech-to-text. Seven discussions were conducted for **RQ5**. Table I lists the statistics of the actual conversational data.

TABLE III
RESULTS OF **RQ2.**

| | TDT | Queue | Memory |
|---|---|---|---|
| Correct ratio | 28.5% | 96.6% | 97.0% |
| Transition analyzing time (sec.) | 5.52e-06 | 0.08 | 1.59 |

*5) Parameters:* The value of $N$ in the cue model is set to 4. We set 0.2 and 0.8 as $\alpha$ and $\beta$, respectively, in the discussion transition analysis.

*6) Baselines:* We compared our algorithm with the following methods:

- TDT: This is a TDT method proposed by [7]. For texts similar to each other collected using cosine similarity, this method selects texts that increase the entropy calculated with entities.
- LDA: We use LDA for only the main theme analysis. Indeed, previous main theme analysis performs LDA [1]; words with the highest weight from the topics of the LDA were assigned as keywords representing the discussion content. In this study, we trained LDA on the W2E dataset for comparison.

*7) Evaluation Criteria:* We calculated the number of times the proposed algorithm determined the appropriate transition of the events of W2E topic to test **RQ1**. For **RQ2** and **RQ3**, we determined the number of times the algorithm identified the manually inserted noise. We manually checked whether the assigned Wikipedia categories were suitable as the main themes for events in the evaluation of **RQ4**. We obtained the **RQ5** results using the same procedure as for **RQ1**$\sim$**RQ4**; however, the authors and three volunteers performs the confirmation processes manually.

### *C. Results on the W2E dataset*

> **RQ1.** Can our algorithm detect properly transitioning discussions?
> **A1.** Both queue and memory models could correctly parse approximately 97% of the text.
> **A2.** The memory model correctly parsed more text than the queue model.

Table III shows the results of the discussion transition analysis of the baseline and proposed models. Results show that the proposed models achieved a higher accuracy (approximately 97%) than the baseline. When we look at the number of incorrect results, the queue model had 50, whereas the memory model had 44, indicating that the memory model was more accurate. In contrast, the queue model took less than 0.1 second to analyze the discussion transition, and the memory model took approximately 1.6 seconds. The queue model counts the number of occurrences by restricting the analysis to the most recent instances. However, the memory

[4] This extended W2E dataset is available from https://onl.tw/jQsLdtP
[5] This extended W2E dataset is available from https://onl.tw/ppekiMM

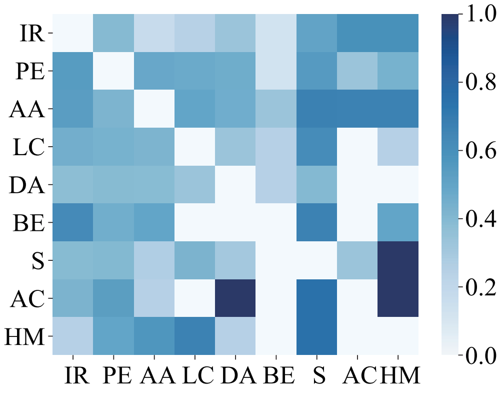

Fig. 3. Category pairs resulted as not perfectly different discussions

model calculates weights exponentially for occurrences in old discussion texts; thus, the computational cost tends to be high. However, this is not a major problem in practice because the analysis was completed within 1 second.

**RQ2.** Can our algorithm detect stalling discussions?
**A.** Our algorithm could detect stalled discussions adequately in 99.6% of the text. The detection failure was due to the failure to retrieve valid entities and Wikipedia articles.

In this evaluation, the queue and memory models accurately detected stalled discussions for 268 of 269 (99.6%) intentionally inserted redundant events. We examined the one event in which the proposed models failed. We examined the nouns, Wikipedia articles, Wikipedia categories, and entities extracted from the correctly analyzed 268 event texts and found their mean values were 10.9, 24.9, 170.1, and 5.4, respectively. In contrast, the mean values for failed detections were 6.0, 10.0, 74.0, and 0.0, respectively. Results show that all the mean values for detection failures were less than the events that were detected correctly. Particularly, no entities could be extracted from small discussions.

**RQ3.** Can our algorithm detect deviating discussions?
**A.** For all texts, we could detect unrelated discussion transitions.

The analysis results for **RQ3** were similar for the memory and queue models. For all 2,055 analyzed events, we could determine that the discussion transition was unrelated. A detailed analysis of the results showed that 1,150 could be analyzed as perfectly different discussions. The remaining 905 events had a little correlation with the previous transition.

Fig. 3 shows the topic categories comprising the 905 events. The categories with particularly high proportions were **S** mixed with **HM** and **AC** with **DA** or **HM**. Table II shows that **S** and **AC** had the lowest average number of events in a single topic (about 3∼4). If this value is low, less information is available for analyzing the discussion transition, and the presence of even a few common entities or Wikipedia categories will have a greater impact. Thus, the proposed models predicted them as not a “perfectly different discussion and analysis” but reported them to have little correlation with the transition.

**RQ4.** Can our algorithm assign the Wikipedia category as the main theme for text?
**A.** The proposed algorithms obtained the appropriate category with 70% accuracy.

For this evaluation, we randomly selected 100 events from the filtered W2E dataset. Subsequently, we manually evaluated the suitability of the results. As this study was designed to support teachers during group learning, we set the criterion for this evaluation as the commonalities between model results and the events. In other words, if the same proper noun was used in the event text, we consider that it is correct. In addition, if the proper noun was the name of a place (e.g., Tokyo) but the Wikipedia category was a country name (e.g., Japan), we also consider that it is correct.

The algorithm accuracy for predicting the main themes from the topics using LDA as a baseline was 32%. The proposed queue and memory models achieved accuracies of 69.3% and 72.7%, respectively. Therefore, it was difficult to apply LDA to determine the main theme, whereas our models performed well in many cases. Next, we examined the analysis time for the two proposed models: the queue and memory models took 4.82 and 1.44e-4 seconds, respectively. The queue model required more computation time than the memory model because it computed the center of gravity of the feature vectors of the Wikipedia categories in the vector space, whereas the memory model used the top of its analysis results. However, the queue model produced results within 5 seconds, which is sufficiently fast not to be a problem considering actual usage.

### *D. Results on actual conversational text*

**RQ5.** How well does our algorithm work for actual conversational texts?
**A1.** For 28 of the 33 conversational texts, Wikipedia categories relevant to the main theme of each session were appropriately obtained.
**A2.** There was only one time occurrence during the conversation where the topic was completely unrelated to the discussion; however, the proposed models could detect it correctly.
**A3.** The models could correctly analyze the discussion transition for 20 of 25 conversational texts.

Finally, we analyzed the performance of the proposed algorithms on actual conversational data. Three Japanese who had completed their graduate studies participated in this discussion. This conversational data contains the actual 7 discussion text data that were collected and separated into one-minute texts. After excluding the conversations with failed speech recognition, 33 actual conversational texts were recorded. A text discussing the differences between birds and dinosaurs, intended for a science unit, is given below. The text is translated from Japanese into English.

*I love birds so much. I'll do some research on Triceratops. Birds, dinosaurs, nobody eats anything or doesn't eat anything, all sorts of things. Yes, there is a difference. It's a difference, isn't it? Yes, yes, yes. Yes, yes, yes. We're talking about a dinosaur that's close to a bird. Uh...*

The Wikipedia article, translated to English from Japanese for the paper presentation, obtained by applying ESA to this text, is as follows: "Lists of extinct species," "Translation of Dinosaurs! - Discover the giants of the prehistoric world," "Dinosaur King," "Torii," "List of birds of Korea," "List of birds of Japan," "Edo o Kiru," "List of Prefectural road of Tottori," "Torii Ryūzō," "Mito Kōmon." From these articles, we extracted the following three categories. "Category:Dinosaurs in video games," "Category:Dinosaurs in comics," "Category:Torii." The obtained categories include "Torii" (鳥居), meaning a gate of a shrine, because it has the same characteristics as the Japanese word for birds (鳥), which is a part of the discussion topic. However, it does not represent a discussion. The other obtained categories were related to dinosaurs, the actual discussion topic.

The following is a continuation of the above conversation: *I see. Okay, okay, okay, okay. What is it again? There's a bird that can't fly. But it seems to me that the story starts with the eating part. But they share the same characteristics as birds that can fly. Yes, yes, yes, yes, yes... I see. That's good. Can I start with the characteristics? First of all, it has feathers on its body... feathers... Feathers...*

We then obtained Wikipedia articles for the Japanese text. "Japan Association of City Mayors," "Feathered dinosaur," "List of birds of Japan," "List of Prefectural road of Tottori," "List of City planning area," "Mito Kōmon," "List of Elementary school in Osaka," "Mito Kōmon." From these articles, we extracted the category "Category:Feathered dinosaurs."

The above results confirmed that Wikipedia categories could be extracted correctly from actual conversational texts. Furthermore, the results for all conversational texts verified that the relevant Wikipedia categories were obtained for 28 of the 33 main themes from each session. In the actual discussion, the texts included conversations that were completely unrelated to the discussion. However, the proposed models could determine that the text was unrelated to the discussion. Therefore, we checked whether our models could detect discussion transition for 25 conversational texts by removing 1 unrelated conversation and the first conversational text for 7 conversations. Our results revealed that the models could correctly analyze the discussion transitions for 20 conversational texts. When we evaluated the discussion transition analysis on actual conversational texts, we obtained approximately 80% accuracy. During the 5 analysis iterations, one wrong result was displayed to the teacher who had to intervene in the discussion. However, this does not interfere with the learning activity, because the teacher could verify that the discussion is not facing any challenges, by analyzing the actual conversation.

## V. Conclusions & Future Work

We proposed algorithms to support teachers in monitoring group learning sessions easily. Based on the assumption that *discussion content changes over time*, the proposed algorithms compare a group's discussion text with previous content. The implemented algorithms assign a Wikipedia category as the main theme of each discussion. We conducted experimental evaluations to confirm the effectiveness of our algorithms and found that the proposed algorithms performed well. Additionally, the proposed queue model provided faster results for discussion transition analysis, whereas the proposed memory model was more accurate. Finally, during the evaluation of allocating the main theme, we found that the memory model provided immediate results, whereas the algorithm finding the center of gravity was more accurate.

Further research is necessary to evaluate the quality of the discussions. The current algorithm assumes that discussions are continuously changing. However, the actual discussion may involve a quality debate that brings together multiple topics that have already been discussed. It is important to evaluate the transition of the discussion by representing it as a tree structure to capture such discussions properly.

## References

[1] Thushari Atapattu and Katrina Falkner. 2016. A Framework for Topic Generation and Labeling from MOOC Discussions. In *L@S '16*. Association for Computing Machinery, New York, NY, USA, 201–204.

[2] Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of Semantic Representation: Dataless Classification. In *AAAI'08*. 830–835.

[3] Hermann Ebbinghaus. 1987. *Memory : a contribution to experimental psychology*. Dover Publications.

[4] Tuan-Anh Hoang, Khoi Duy Vo, and Wolfgang Nejdl. 2018. W2E: A Worldwide-Event Benchmark Dataset for Topic Detection and Tracking. In *CIKM '18*. Association for Computing Machinery, 1847–1850.

[5] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *ICML'14*. JMLR.org, II–1188–II–1196.

[6] Sports Science Ministry of Education, Culture and Technology. 2018. Koutou Gakkou Gakushuu Sidou Youryou Kaisetsu Chiri Rekishi Hen. http://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2018/08/29/1407073_03_1.pdf

[7] Kira Radinsky and Sagie Davidovich. 2012. Learning to Predict from Textual Data. *J. Artif. Int. Res.* 45, 1 (2012), 641–684.

[8] Taoufiq Zarra, Raddouane Chiheb, Rdouan Faizi, and Abdellatif El Afia. 2018. Student Interactions in Online Discussion Forums: Visual Analysis with LDA Topic Models. In *LOPAL '18*. Association for Computing Machinery, Article 30, 5 pages.