Journal Title
XX(X):1–14


DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE


# Real-time Discussion Transition and Main Theme Analyses for Teacher Support

**Yasunobu Sumikawa[1] Akikazu Takada[2] Akira Ichinose[3] Ari Murakami[2] Yuki Toyono[2] Ryohei Ikejiri[5] Meiko Sakasegawa[5] Kaoru Sekine[6] Yuhei Yamauchi[5]**


### Abstract

The effectiveness of group and online learning environments has been widely recognized. Hence, supporting teachers monitoring of group activities in online learning environments is increasingly important. In this study, we propose algorithms for analyzing discussion transitions and their main themes. The first algorithm aims to help teachers quickly identify groups that require intervention. It uses entities and Wikipedia data in the discussion texts to create models that capture continuously changing discussion content over time. The second algorithm aims to instantaneously identify the discussion topic of the group prior to teacher intervention. From the Wikipedia categories obtained regarding discussion transition analysis, the algorithm selects those with the highest probability of being appropriate for the 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 discussion dataset. The results confirmed that the algorithms detected the discussion transitions and main themes with high accuracies.


### Keywords

Discussion analysis, Transition analysis, Topic detection and tracking, Wikipedia


## 1 Introduction

The effectiveness of collaborative learning in groups as a high-dimensional learning approach has been widely recognized. The Ministry of Education, Culture, Sports, Science and Technology (2018) has recognized the effectiveness of collaborative learning as a high-dimensional approach within its new guidelines. This aligns with broader trends in academic fields like educational technology and science learning, as well as frameworks established by the Organisation for Economic Co-operation and Development.

Because computers have become widely used, attempts have been made towards leveraging their convenience in the field of learning. As the Internet continues to grow, interactive classes are now conducted over the Internet using tools that allow online communication. When group learning is conducted using such a communication tool, other members are blocked using a function that allows only designated members to work as a group. Although such tools are important features for group learning, teachers struggle to monitor the learning activities and manage simultaneously all groups because of the high functionality of these partitions. That is, changes in the learning environment increase the need for teacher support.

This study aims to support to monitor and interventions by teachers in group learning. For this purpose, we propose unsupervised learning algorithms that analyze 1) whether groups can conduct appropriate discussions and 2) the discussion topics of the groups to ease intervention. The following provides an example discussion requiring teacher intervention. In this example, we consider that two students, A and B, are discussing the difference between gasoline and electric vehicle (EV) cars.

A: Let's find out the differences between gasoline and EV cars.
B: Sure. I will look into gasoline cars.
A: Alright, in that case, I will look into EV cars.
A: Look at this document. It describes the different types of motors that EVs use.
B: Good find! Is there any difference in the way the fuel is refilled?
A: Let's find out how fuel is replenished now.
B: Speaking of which, what are you going to do after school today?
A: I haven't decided yet.
B: Oh, an EV car needs 6 hours to charge!
A: Really? Why does it take so long? A gasoline car would require only a few minutes.
B: Maybe we need to know something about chemistry to explain the charging rate, but I do not know anything about it.
A: Me neither.

[1]Takushoku University, Japan
[2]Ddrive K.K., Japan
[3]First Torrent LLC., Japan
[4]The University of Tokyo, Japan
[5]Tohoku University, Japan

**Corresponding author:**
Yasunobu Sumikawa, Takushoku University 815-1 Tatemachi, Hachioji-shi, Tokyo, Japan.
Email: ysumikaw@cs.takushoku-u.ac.jp

In the example conversation, the students appropriately discussed with reference to the document until they examined the fueling methods. However, after deciding to conduct research independently, student B remarked: *What are you going to do after school today?* This conversation should not occur during class. Subsequently, B's discovery of the fueling features of EV vehicles steered the conversation back to the original topic. To explain these features, A and B realized that they required chemistry knowledge. However, their lack of chemistry knowledge made continuing the conversations difficult. In this situation, our algorithms can analyze two points in this group's conversation regarding their inappropriate discussion to then convey these results to the teacher. Note that we assumed that the students immediately resumed a discussion that is properly suited to the class content, even if they discussed their afterschool activities, which is irrelevant to class content. However, if students discuss topics unrelated to the learning content for long periods of time, the teacher should intervene to avoid wasting learning time.

The main questions that our study aims to answer are as follows:

1. *Can our algorithm detect properly transitioning discussions?*
2. *Can our algorithm detect stalling discussions?*
3. *Can our algorithm detect deviating discussions?*
4. *Can our algorithm find the main themes for texts?*
5. *How well does our algorithm work for actual conversational texts?*

We investigate these questions by proposing algorithms for discussion transitions and main theme analyses. The transition analysis algorithms assume that *the content of the discussion changes over time without gaps.* To capture this assumption, we propose two models: The first is called the queue model, and it limits the scope of the analysis to the latest texts. Second, the memory model distinguishes between recurring and passing topics. The results of the transition analysis can identify groups requiring teacher intervention such that the discussion topics can be discussed appropriately. At the time of intervention, our algorithm also analyzes the main theme of the discussion to immediately know what the group has discussed thus far. We propose two models for analyzing of the main theme. First, we map Wikipedia categories onto a coordinate space and consider the category closest to the center of gravity as the approximate solution. Second, we consider the topic with the highest frequency in the memory model analysis results as the main theme.

Our algorithm applies the aforementioned models when it receives a text representation of the discussion. The results obtained by applying the two models are a numerical value representing the appropriateness of the discussion transition and text representing the main theme. The visualization of these results, which can be displayed by group, provide support to teachers.

To evaluate the effectiveness of our models, we conducted a quantitative evaluation using the W2E dataset and a qualitative evaluation using elementary school students' discussion texts collected in actual elementary school classes. The W2E dataset consists of topics, including news articles regarding a single event reported by various news agencies. We simulated the progression of discussions by considering news agencies as engaging in discourse on each topic through the publication of newspaper articles. The evaluation results demonstrate that our models correctly analyzed the discussion transitions for approximately 97% of the text. We then created subsets from the W2E dataset to simulate deviating and stalling discussions and found that the proposed models could correctly detect almost all of them. The models allocated appropriate Wikipedia categories as the main themes for discussion texts, with an accuracy of approximately 70%. We also found that our models performed good analyses on the actual discussion texts because our model achieved a 50% accuracy in this evaluation.

Sumikawa et al. (2022) provided the basis for this study in a paper presented at the WI-IAT 2022 conference. Compared with the previous study, this study generalizes the 1) input text assumed by the proposed models to more closely resemble the actual teaching style and 2) equations of our models to use the input text. In our new models, external information resources are added to incorporate any information regarding the discussion contents into the algorithm. This allows the analysis of the learner's texts to check if it is in line with the intended learning of the group. We performed experimental evaluation using the new algorithms. In addition, this study evaluated the effectiveness of the proposed algorithms using actual learners' conversational texts. The previous study evaluated the effectiveness of the algorithm based on discussions between adults who had completed graduate education. The conversational texts used in this study are more appropriate for validation in an actual classroom setting. Finally, this study discusses the differences between our study and previous studies that used machine learning (ML) to support learning more extensively than the previous study. This facilitates clear demonstrations of the novelty of our study and of how the proposed algorithms can be combined with methods proposed in several previous studies to develop classes.

The remainder of this paper is organized as follows. In the following section, we present related work. Section 3 describes the proposed algorithms. Section 4 presents experimental evaluation results and detailed analyses. Finally, the last Section concludes the paper and outlines directions for future work.

## 2 Related Work

This section surveys studies in which topics are analyzed to make sense of text content, ML is used to analyze group learning, and related texts is collected in time series. First, we examine a study on the aspects discussed from sentences about a particular topic (Section 2.1) and a study on topic transitions (Section 2.2). We then describe literature proposing ML algorithms to support learning activities in Section 2.3. Finally, survey topic detection and tracking (TDT) that determines whether newly input text is the same as previously accumulated text (Section 2.4).

### 2.1 Topic Aspect Analysis

A single discussion content may involve several different events, objects, persons, and so on. Within the fields of natural language processing and ML, Carbonell and Goldstein (1998); Singh et al. (2016) and Chatterjee and Dietz (2021) have extensively explored diversity techniques for retrieving various aspects and events from given text.

Identifying the meaning of a word or phrase in a sentence may be classified into this type of research. Raganato et al. (2017) proposed word sense disambiguation to involve uniquely determining such meaning. In addition, named entity recognition disambiguation (NERD) is being actively studied that not only uniquely determines the meaning, but also detects named entities in the text. Several NERD tools have been developed to facilitate entity extraction and disambiguation by processing input sentences and linking them to Wikipedia articles. These include TagMe introduced by Ferragina and Scaiella (2010), DBpedia spotlight presented by Mendes et al. (2011), and AIDA developed by Hoffart et al. (2011). For analyzing the meaning of words, Nanni et al. (2018) proposed a method for extracting aspects related to a given entity from Wikipedia articles. In addition to traditional Wikipedia-based research, recent studies have expanded to other text types. For example, Prasojo et al. (2015) performed studies on comment sentences, while Shen et al. (2010) focused on the analysis of query sentences. To evaluate the effectiveness of these methods, Ramsdell and Dietz (2020) constructed a ground-truth data set.

All the aforementioned studies aimed to improve user convenience. In the present study, we develop algorithms for supporting teacher monitoring; thus, our algorithms and the methods of the previous studies do not overlap in function. None of the aforementioned methods focused on learning activities; however, if the reported methods are used for group learning, they can be useful when learners seek knowledge about a deep and broad range of topics.

### 2.2 Learning Content Analysis & Visualization

Since massive open online courses have become widely used, many studies have been conducted to analyze learning activity in online learning environments. ML techniques, specifically text classification as explored by Zhou et al. (2020) and Mulla and Shaikh (2024), are employed in systems designed to support learning activities. In these studies, ML is used to analyze information such as the learner's progress and level of understanding. The analysis targets of previous studies can be categorized as follows: 1) analysis of learners' interactions as a network, 2) analysis of learners' emotions, 3) analysis of whether learners have newly acquired knowledge, 4) topic measurement, 5) analysis of speech and questioning behavior, and 6) analysis of learning patterns. All of the above are useful to assist teachers in developing better lessons by helping them understand the learners' situation. In the remainder of this sub-section, we present topic measurement in detail, because it is related to the present study. Detailed surveys of the other five types of studies are provided by Zhang et al. (2022b) and Ahmad et al. (2022).

Topic measurement is useful for teachers monitoring learning activities because it automatically analyzes the discussion content. Several studies have been performed on topic measurement, mainly using latent Dirichlet allocation (LDA) proposed by Blei et al. (2003). LDA is an unsupervised learning technique that analyzes word combinations in a sentence to reveal the word clusters that constitute a topic. Chen (2014), Zarra et al. (2018), and Hsiao and Awasthi (2015) applied LDA to discussion texts to visualize the word groupings used within those conversations. In one study, Ezen-Can et al. (2015) applied clustering to Bag-of-Words (BoWs) instead of LDA; however, this method allows only words that appear in the text to be topics.

This paper also proposes a topic measurement method; however, instead of assigning the words used for training as topics as in the aforementioned studies, our algorithm allocates Wikipedia categories. Therefore, our algorithm can use words that are not directly used in the discussion as topics. Even if there is no appropriate word in the discussion text to be used as a topic, the appropriate topic can be determined from the vast data of Wikipedia.

Many visualization systems using LDA aim to assist teachers in giving better feedback to learners. As the previously reported methods and our algorithms do not interfere with each other, teachers can use them simultaneously for better classroom management. For example, from the beginning to the middle of the class, the visualization systems of previous studies can be used to understand what type of content the learners are discussing. If the discussion is proceeding properly, the teacher will guide it in a better direction. In contrast, if a discussion has not been initiated or cannot be continued, the teacher can quickly find the students using the proposed algorithm and give them a hint after reviewing the conversation thus far.

### 2.3 Learning Support Environment using Machine Learning Techniques

In contrast to studies addressing out-of-school learning examined by Ashikawa et al. (2019), this study focuses on in-class learning. Accordingly, we discuss the differences between our work and existing classroom-based learning support systems. In recent years, research on supporting learning activities through Large Language Models (LLMs) has advanced rapidly. Specific areas of investigation include the potential of LLMs to function as tutors, as explored by Lieb and Goel (2024) and Shochcho et al. (2025). Additionally, Wang et al. (2025b) examined LLM-based support for personalized learning within online environments, while Wang et al. (2025a) developed frameworks utilizing LLMs to facilitate personalized, goal-oriented learning. Research by Yang et al. (2025) has also addressed learner support for learning-oriented search tasks. Collectively, these studies primarily aim to enhance direct learner support.

Despite the proliferation of studies targeting learner and teacher support and the emergence of functional frameworks for LLMs proposed by Chu et al. (2025), the detection of discussion stalls in group learning remains unexplored. To

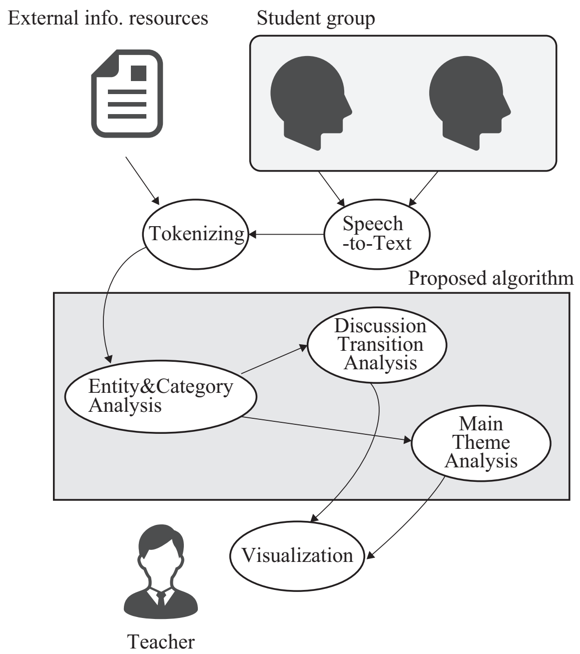


**Figure 1.** System overview

the best of our knowledge, this study is the first to address this specific challenge.

## 2.4 Topic Detection and Tracking

Finally, we survey TDT, a ML study that follows the same procedure as the discussion transition analysis of this study. In both TDT and this study, analysis is performed to determine whether the content of the newly input text is identical to the content of the previously accumulated text. In TDT, this procedure is often applied to newly reported news and Tweet text to generate a sequence of events arranged in chronological order. Radinsky and Davidovich (2012) proposed a method for generating chains of past events to predict future events. Within the field of TDT, Qi et al. (2017) developed methods for detecting and tracking hot events from online news streams. Similarly, Tan et al. (2014) introduced mechanisms using local community detection to distinguish between global and local hot events. While previous studies created linear timelines for event representation, recent researchers have sought to generalize event dependency relationships through more complex structures. Specifically, Zhang et al. (2022a) introduced the Storytree for tree-based representations, while Liu et al. (2020) proposed the Story Forest. Additionally, Sawahata et al. (2026) and Machizawa et al. (2026) developed StoryNetworks to represent these relationships via network structures, and Yang et al. (2009) utilized the event evolution graph (EEG) to model them as graphs.

These TDT methods are similar to this study in procedure; however, they are not designed to support learning in the first place. It is difficult to use the TDT methodology because there is no mechanism to present useful information to teachers monitoring learning activities.

# 3 Algorithms

An overview of the proposed algorithms is presented in Fig. 1. The algorithms use two types of documents: discussion texts and external information resources. The former is derived from the discussion of the learners. The latter refers to any document, for example, a textbook, reference book, Wikipedia article describing the content to be studied in the discussion, or a list of keywords that learners should pay particular attention to. In the absence of appropriate external information resources, we consider this document as an empty string. After tokenizing the two types of documents, we extract entities and Wikipedia categories from them that are used in discussion transition and main theme analyses. Finally, we present the results of these analyses to the teachers.

We describe the processes of extracting entities and Wikipedia categories from discussion texts and the discussion transition and main theme analyses in the following subsections.

## 3.1 Entity & Category Analyses

Our algorithms use entity and Wikipedia categories to represent the actual words used in the documents and their meanings, respectively. Because we use the entity extraction results, any extraction algorithm is suitable[*]. Additionally, explicit semantic analysis (ESA) proposed by Chang et al. (2008)[†] is used to obtain the Wikipedia categories. ESA calculates the relevance of given texts and each Wikipedia article and then ranks the articles based on their relevance scores. After retrieving Wikipedia articles, we extract their Wikipedia categories assigned by Wikipedia editors. These annotations were performed by manually checking whether the categories represent the articles. Thus, our algorithm uses the Wikipedia categories as abstractions of the Wikipedia articles. This allows us to analyze the themes of the input discussion texts. To improve the analysis results, we remove unrelated Wikipedia categories by checking whether they contain nouns used in the discussion text. We combine the entities and Wikipedia categories as a list *words* for the following analyses.

## 3.2 Discussion Transition Analysis

In this analysis, we assume that the discussion contents should change over time without any gaps. To implement this assumption, we design two discussion transition models: queue and memory models. The former focuses on the latest texts, whereas the latter focuses on repeated texts.

*3.2.1 Queue model* This model analyzes only the most recently generated texts by excluding old ones using a queue, which is the first-in-first-out order data structure. If a discussion text is inputted, the model stores the text in the queue and removes the oldest text. To store only the latest

[*] Our implementation performs AutoML as an entity extraction. https://cloud.google.com/natural-language/automl/docs

[†] We used Descartes in our implementation that are available at https://cogcomp.seas.upenn.edu/page/software_view/Descartes

$N$ appropriate discussion texts in this queue, this model first analyzes whether the input text should be stored.

The analysis algorithm is as follows. We have used three symbols to describe the algorithm: $input$, $qtext$, and $econt$ as the input text, texts in the queue, and external information resources, respectively. In this analysis, we examine $input$ to determine whether it is similar to $qtext$ and $econt$. First, all texts in the queue are loaded for each analysis. After creating all texts $AllText$ by combining $qtext$ and $econt$, we create BoWs, named $AllW$, for $AllText$. Subsequently, we create BoWs for the $input$ to calculate similarity from $AllW$. As the similarity score $score$ can be expressed in a range between 0.0 and 1.0, the appropriateness of the discussion is determined by whether this score is within the specified range. The formal equation is $\alpha < score < \beta$. If the $score$ is below $\alpha$, the text is regarded as deviating because the $input$ and $AllText$ do not share common words or topics. If the $score$ is higher than $\beta$, the text is regarded as stalling because new words or topics were not used in the input text.

We present the formal definition of this model as follows:

$$QVal(input_W, AllW) = \frac{\mid input_W \cap AllW \mid}{\mid input_W \cup AllW \mid} \quad (1)$$

where $input_W$ is words in the $input$. The numerator represents the number of words commonly shared by $input$, $qtext$ and $econt$. The denominator is the number of words used in the texts.

*3.2.2 Memory model* An intuitive explanation of this model is that *although the discussion changes over time, repeatedly appearing text should become important topics of the discussion*. To implement this idea, we utilize the memory forgetting curve developed by Ebbinghaus (1987), which illustrates how information retention declines over time and improves through repeated learning.

An example of the memory model calculating the importance scores of words is shown in Fig. 2. The example involves the calculation of weights of three words `Soccer`, `Japan`, and `Sports` using the model. These words are included in the texts generated at times `t'`, `t''`, and `t'''`. Because this model regards repeated words as important, it computes a weight for each occurrence of a word. These weights are reduced over time; however, the final result is the sum of the weights at each point in time. Thus, words that occur only a few times have their weights reduced, whereas those that occur repeatedly have relatively large weights owing to the addition of the weights.

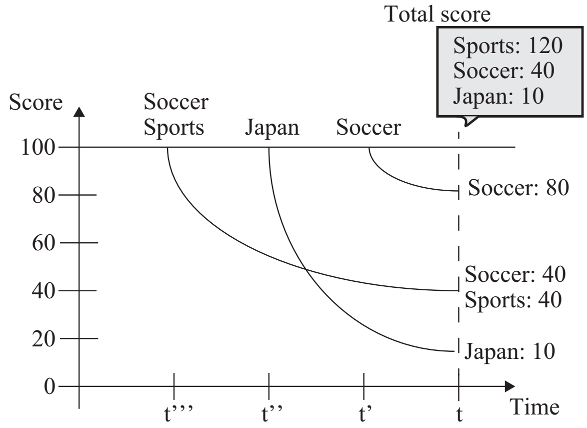


**Figure 2.** Example for memory model

The memory model reduces the weights by applying the exponential function $e$. To represent the passage of time, we set $t - t'$ as an argument of the exponential function, where $t$ represents the time at which the text is being analyzed and $t'$ represents the time at which the text is generated. While $t$ is updated each time the weight is calculated, $t'$ is invariant. The simple computation of these differences reflects the passage of time.

As the memory model calculates weights for words in all texts, it is useful to have large weights for texts that are relevant to the learning content and small weights for the remaining texts. To express this idea, we introduce the parameter $\omega$. If the generated text is similar to the documents of external information resources, this parameter causes the weights to decrease slowly. However, if it is not similar, the time difference is increased. The weight reduction is given as follows:

$$f(t, t', \omega) = \omega \times e^{(-(t-t'))}$$

We next define the equation for calculating the weight for each word $w$ as follows:

$$W(w, t, \omega) = \sum_{t' \in times(t)} f(t, t', \omega) \times \delta(w, Words(t'))$$

where the function $Words(t)$ loads the words of a text generated at time $t$. The function $times(t)$ retrieves the time recorded before time $t$. $\delta$ is a function that returns 1 if $w$ is used in $Words(t')$; otherwise, it returns 0.

Finally, we define the formal equation for the sum of the weights as follows:

$$MVal(input_W, t, \omega) = \sum_{w \in input_W} W(w, t, \omega) \quad (2)$$

After calculating the weights, the memory model determines whether the $input$ is related to the discussion. This is achieved by comparing the results of Eq. 2 with two thresholds, $\alpha$ and $\beta$, as well as with the queue model.

### 3.3 Main Theme Analysis

The main theme analysis allocates a Wikipedia category to each discussion text. As described in Section 3.1, we obtain Wikipedia categories at the first processes in the proposed algorithm. Our main theme analysis treats Wikipedia categories as summaries of articles because they are shorter than the article text and are deemed relevant by Wikipedia editors. As we obtain the Wikipedia categories after retrieving Wikipedia articles by applying ESA to the discussion text, we regard the categories as an abstraction of the discussion. Therefore, we select the appropriate Wikipedia category as the main theme. We propose appropriate selection methods for each of the two discussion transition analyses.

*3.3.1 Queue model* As the queue model treats the weights of each Wikipedia category uniformly, we embed all categories in the coordinate space to analyze the main theme. This is done using Doc2Vec developed by Le and Mikolov (2014), which is a widely used method for embedding. As

**Algorithm 1** Algorithm overview

**Input:** An input text $input$, past discussion texts $N$, texts from external information resources $EIR$, thresholds $\alpha, \beta$
**Output:** A score $score$, main theme $m$

1: **Function** $DiscussTrans(input, N, EIR, \alpha, \beta)$
2: // Text Analysis
3: $\omega \leftarrow Sim(input, EIR)$
4: $input_W \leftarrow ExtractEntityWikiCats(input)$
5: $AllW \leftarrow ExtractEntityWikiCats(N, EIR)$
6: // Discussion Transition Analysis
7: $score \leftarrow$ a result of Eq. 1 or Eq. 2
8: **if** $\alpha < score < \beta$
9: $N \leftarrow input_W$ // Store $input_W$ to discussion texts
10: **end if**
11: // Main Theme Analysis
12: $m \leftarrow$ results of Section 3.3
13: **return** $score, m$

Doc2Vec requires description texts, we collect descriptions of these categories. In general, each category has one or more Wikipedia articles. We regard the texts of the articles as category descriptions. Once we train the Doc2Vec on the collected texts and apply it to all categories, we obtain feature vectors for the categories. Finally, we set the center of the discussion text using the collected Wikipedia categories. The queue model selects the category closest to the center of gravity of the categories obtained for each discussion text as the main theme.

*3.3.2 Memory model* The idea behind our memory model is to ascertain important Wikipedia category. The memory model considers that the higher the frequency, the more important is it. Therefore, this model simply selects the category with the highest score in Eq. 2 as the main theme.

### 3.4 Algorithm Overview

The pseudocode for our algorithm is presented in Algorithm 1. Initially, the algorithm calculates the similarity between the input and external information resources texts to set the value of $\omega$ for the memory model. We then extract entities and Wikipedia categories from all texts, as shown in lines 4$\sim$5. After these preparations, the discussion transition analysis is performed by applying Eq. 1 or Eq. 2. Lines 7$\sim$10 indicate that the discussion transition analysis decides if the input text properly represents the progress. Finally, the algorithm obtains the main theme for the input text and returns the outputs of the two analyses.

## 4 Evaluation

### 4.1 Experimental Settings

*4.1.1 Datasets* In this study, we conducted four quantitative evaluations and one qualitative evaluation. For the quantitative portion, we utilized the W2E dataset, which Hoang et al. (2018) originally developed for the TDT problem. The W2E dataset organizes related newspaper articles into topics, with each article sourced from different news agencies. This study focuses on this aspect and regards each topic as a collection of texts that discuss the analytical results of the same event as reported by various news agencies. Table 1 presents the examples from this dataset. The W2E dataset creator manually collected and aggregated news from the English version of Wikipedia to create the topics. The dataset contains 3,083 topics reported in 2016. Although the W2E dataset is designed for the TDT problem, some topics include only one or two events. For our study, we filtered topics by checking whether the number of events in a topic was greater than 3[‡]. Consequently, we used 1,781 events and 269 topics. Table 2 shows the statistics of our dataset. These topics were classified into 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**). The statistics for each category of the filtered W2E dataset are presented in Table 3.

The filtered W2E dataset was extended to create duplicated events in a topic and to combine the two topics to evaluate the ability of our algorithm to detect stalling and deviating discussions. During the duplication process, we duplicated the last news text for each topic. This dataset contained 269 topics, as well as a discussion transition analysis[§]. To combine the two topics to simulate deviation, we selected two unrelated topics. We revealed two unrelated topics through two examinations. First, we simply removed combinations if their news articles shared nouns. Second, we checked their dependencies; we removed combinations if one topic depended on another. To check the dependencies, we created feature vectors using latent semantic analysis introduced by Deerwester et al. (1990). Next, we applied the mutual information (MI) to all the event combinations. The MI is defined as follows:

$$MI(A, B) = \sum_{a \in A} \sum_{b \in B} p(a, b) \log \left( \frac{p(a, b)}{p(a)p(b)} \right) \quad (3)$$

A high value indicates a high degree of dependence. If this value was greater than 0.3, the combination was considered dependent and discarded. As a result, 2,055 topics were generated[¶].

For qualitative evaluation, we collected discussion texts from actual elementary school classes. We prepared learning content as external information resources that allowed students to learn mathematics and social science in a discussion format to match the content of the class on the day of the experiment. All the students were Japanese speakers. Before conducting the experiment, the students and their parents were asked to fill out a consent form for cooperation in the experiment. Agreement was obtained from all the students and their parents. We used our system in two classes: mathematics and social science. First, we collected speech data by using Google Cloud Speech-to-Text[‖] during the mathematics class for parameter adjustment of our algorithm. We then used the algorithm in the social science

[‡] The filtered W2E dataset is available: https://onl.tw/NnFiEXM
[§] This extended W2E dataset is available at https://onl.tw/jQsLdtP
[¶] This extended W2E dataset is available from https://onl.tw/ppekiMM
[‖] https://cloud.google.com/speech-to-text

**Table 1.** Example of test dataset. The first column represents topic categories of the second and forth columns. The second and forth columns are the topic ID of the third and fifth columns' topics in W2E dataset.

| Cat. | TOPIC ID | Event texts | TOPIC ID | Event texts |
|---|---|---|---|---|
| **PE** | TOPIC-896 | 2016-07-14 Elizabeth Truss is named Secretary of State for Justice and first ever female Lord Chancellor of the United Kingdom as former chancellor Michael Gove is ousted from the cabinet.<br>2016-07-13 The new Prime Minister of the United Kingdom Theresa May begins forming her ministry following the end of the Second Cameron ministry. | TOPIC-1780 | 2016-06-29 The process to elect a new leader of the Conservative Party to replace outgoing Prime Minister David Cameron begins in the United Kingdom.<br>2016-07-11 Prime Minister David Cameron announces he will step down on Wednesday, July 13.<br>2016-06-30 Former Mayor of London Boris Johnson rules himself out of running in the Tory leadership contest, a move believed to be influenced by Michael Gove's announcement earlier in the day to run for the leadership.<br>2016-07-05 Home Secretary Theresa May gets 165 votes after the first ballot of Conservative members of parliament to select a new Leader and the next Prime Minister. |
| **S** | TOPIC-1534 | 2016-09-10 In tennis, German Angelique Kerber defeats Czech Karolína Plíšková in three sets to win the 2016 US Open women's singles title.<br>2016-09-11 In tennis, Swiss Stan Wawrinka defeats Serbian Novak Djokovic in four sets to claim the 2016 US Open men's | TOPIC-1996 | 2016-01-31 In tennis, defending champion Novak Djokovic of Serbia defeats second seed Andy Murray of the United Kingdom in the men's singles in straight sets. It is Djokovic's third straight Grand Slam title.<br>2016-01-30 In tennis, Angelique Kerber of Germany tops (2–1) defending champion Serena Williams of the United States, 6-4, 3–6, 6–4, to win the Women's Singles. This is Kerber's first Grand Slam title. |

class to determine whether each group discussed the learning theme appropriately. The statistics of the data collected during this evaluation are presented in Table 2. We analyzed whether the proposed algorithms produced appropriate results for the actual conversational data. The subject of the social science class was finding differences between cars of the past and present. This class was conducted for a total of 20 minutes: approximately 6 minutes of discussion in each group, approximately 9 minutes of teacher review of the statuses of all groups, and approximately 5 minutes of discussion in each group. Each group consisted of two students. During the discussion, the students placed a laptop on their desk, and voice data were acquired from the laptop and stored on a server as discussion data. We utilized the Web Audio API in JavaScript to immediately convert audio input from a laptop microphone into WAV format. The resulting audio data was then processed using Google Cloud Speech-to-Text** for transcription. The transcribed text was subsequently transmitted to the server running the proposed algorithms. Since the conversion from audio acquisition to text transcription is performed in real-time, the proposed algorithms can be applied without delay. The acquired discussion data were divided into 1-minute increments so that the teacher could review the discussion minute-by-minute.

*4.1.2 Parameter Settings* We set different parameter values for the two experiments. For the experiments on the W2E dataset, we randomly selected 20 topics from the dataset as development data and used them for parameter tuning. We set $N$ as 4 in the queue model and set $\alpha$ and $\beta$ as 0.2 and 0.8, respectively, in the discussion transition analysis. These development data were excluded from the evaluation data.

In the experiment involving the elementary school, we manually analyzed discussion data obtained in the mathematics class. Values of 0.02 and 0.3 for $\alpha$ and $\beta$, respectively, were found to be appropriate in the discussion transition analysis.

*4.1.3 Baselines* In this study, two methods were used as baselines. The first one was a TDT method proposed by Radinsky and Davidovich (2012). We evaluated this method in the discussion transition analysis for the W2E dataset because the process of this method is similar to our transition analysis as described in Section 2.4. For the main theme analysis, we used LDA as the baseline. As mentioned in Section 2.2, LDA was used for detecting the main theme in previous studies. The visualization system proposed in Atapattu and Falkner (2016) shows words with the largest

**The demonstration is available at https://www.google.com/intl/ja/chrome/demos/speech.html

**Table 2.** Statistics of test dataset.

| | | |
|---|---|---|
| W2E | Num. of topics | 269 |
| | Num. of events | 1,781 |
| | Ave. num. of tokens | 32.9 |
| | Ave. num. of events per topic | 6.62 |
| Elementary school | Num. of subjects for actual discussion | 1 |
| | Num. of discussion texts for actual discussion | 20 |

**Table 3.** Statistics of test dataset by category.

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

weights from the LDA topics. In this study, we trained LDA on the W2E dataset.

*4.1.4 Evaluation Criteria* For the discussion transition analysis evaluation with the W2E dataset, we counted 1) the number of scores between $\alpha$ and $\beta$ that the algorithms output as appropriate transitions, 2) the number of stalling instances found that were created by the duplication, and 3) the number of deviations found that were created by mixing two topics.

For the main theme analysis, we manually inspected each topic to ensure that it was assigned an appropriate main theme by our algorithm and LDA. We used 100 events that were randomly selected for this evaluation from the dataset used in the discussion transition analysis. The objective of this analysis is a teacher support; thus, we regarded the main theme as correct if there were any common words or city-country relationships between the analyzed text and the allocated Wikipedia category. For example, if our algorithm allocated "Japan" for an event that occurred in Tokyo, we regarded the result as correct.

For the actual discussion texts used by elementary school students, we first manually identified the discussion status (transitioning, stalling, or deviating) for each text. We then applied our algorithms to the text and evaluated the numbers that were the same as human judgment.

## 4.2 Answers to the First RQ: Can our algorithm detect properly transitioning discussions?

We first present results of the discussion transition analysis with the W2E dataset for answering the research question introduced in Section 1. Our queue and memory models obtained correct answers for approximately 97% of the text; however, the memory model was better than the queue model. The reasons for this are discussed below.

We first calculated the percentages of events correctly predicted as appropriate for the appropriate transitional discussions by the three methods: TDT, the queue model, and the memory model. The results were 29.0%, 96.7%, and 97.1%, respectively. Thus, the memory model was the most accurate, and TDT was the least accurate. Although the memory model achieved the highest accuracy, the queue model also achieved an accuracy of approximately 97%; thus, both the proposed models achieved high accuracies.

Next, we measured the computation times of the three methods. The average amounts of time taken to complete the analysis for each topic by TDT, the queue model, and the memory model were 6.64e-06, 0.47, and 1.92 seconds, respectively. TDT achieved the shortest time, and the memory model had the longest time. The proposed models are computationally expensive because they involve acquiring Wikipedia data and updating values each time text is entered. However, this is not a problem when the models are used in practice, as the models produced results in 1$\sim$2 seconds. The results indicate that the proposed models are superior to TDT for discussion transition analysis.

Next, we analyzed the categories in which the two proposed models misjudged the input text as inappropriate. Table 4 presents the results. The two models tended to misinterpret the input text in the same categories. In addition, the queue model made misjudgments in the three categories of **AA**, **LC**, and **PE**. When we analyzed which texts were misjudged, we found that if the memory model made a mistake, the queue model also made a mistake. We then focused on the texts for which only the queue model made mistakes and found that the queue model misjudged the 8th, 15th, and 20th texts for three different topics. In this study, the queue model only analyzed the past 4 texts. To obtain correct results for the three texts mentioned above where errors occurred, the queue size needed to be larger than 4. We also analyzed texts misjudged by the memory model and found that it failed to extract important nouns. As an example, we show a topic including the following two events: 1) *The U.S. House Committee on Foreign Affairs unanimously passes a resolution reaffirming the Taiwan Relations Act and the Six Assurances* and 2) *Taiwanese president Tsai Ing-wen refutes accusations that she would have interpreted the 1992 consensus differently.* Our models produced errors for the second event. The nouns extracted for these events were *house, committee, affairs, resolution, relations, act, assurances* and *president, refutes, accusations, consensus*, respectively. "Taiwan" was a common word for these events; however, it was not extracted from either text. This was the reason for the misprediction.

**Table 4.** Misprediction ratios for event categories

| | S | AA | BE | AC | LC | PE | IR | DA | HM | *Total* |
|---|---|---|---|---|---|---|---|---|---|---|
| Queue | 2.2% | 2.5% | 4.5% | 0.0% | 6.7% | 4.0% | 3.1% | 3.7% | 0.0% | 3.3% |
| Memory | 2.2% | 1.8% | 4.5% | 0.0% | 5.6% | 3.6% | 3.1% | 3.7% | 0.0% | 2.9% |

## 4.3 Answers to the Second RQ: Can our algorithm detect stalling discussions?

Next, we evaluated the number of detected stalling discussions for the W2E dataset. Both the queue and memory models correctly detected stalled discussions for 268 of the 269 (99.6%) topics.

We checked the entity and Wikipedia data extracted from the text for which our models failed to detect stalling and found that there were no entities extracted from the text. To investigate this, we examined the average numbers of nouns, Wikipedia articles, Wikipedia categories, and entities for all the texts. First, we compared the average numbers of entities between correct texts and the failed text. The values were 5.4 and 0.0, respectively. These results indicated that the failed text did not use any entities in the text. We checked the average numbers of Wikipedia articles collected from the correct and failed texts and found that the values were 24.9 and 10.0, respectively. As the average numbers for nouns were 10.9 and 6.0, all the numbers of the failed text were smaller than those of the correct text. Thus, the reason for the failure of our models was the insufficient numbers of collected entities and Wikipedia articles.

Next, we performed detailed investigation of events for which our models failed to detect stalling. The description of one event was *The Syrian cessation of hostilities truce is in effect, as of midnight, Saturday, local Syrian time.* This topic included four events, all of which occurred in Syria. The events were duplicated; thus, the topic included five events of four different types. Among the entities and Wikipedia categories extracted for the first three events, there were none in common with the above event. Therefore, our models incorrectly judge the fourth event and the intentionally inserted event as unrelated stories that appeared suddenly.

## 4.4 Answers to the Third RQ: Can our algorithm detect deviating discussions?

In this evaluation, we found that our two models perfectly detected deviating discussions; the models regarded all analyzed events as unrelated for all the texts we inserted.

As we inserted irrelevant text for this evaluation, we analyzed in detail similarities between the inserted text and the texts that preceded it. For 1,150 of the 2,055 topics, our models determined that the intentionally inserted text was completely dissimilar to the previous text; i.e., their similarity scores were 0.0 according to our model. Next, we analyzed the remaining 905 topics that were deemed to have low similarity (its score was greater than 0 and less than 0.2) to determine what combination of categories they were and the results are presented in in Table 5. The vertical axis indicates the category of the event mixed with other topics, and the horizontal axis indicates the category of the topic to which the event belonged.

We can see that all the topics of **AC** mixed into **DA** and **HM** had a few commonalities. In addition, **S** had a commonality with **HM**. Looking at Table 3, among the categories, we can see that the average number of events per topic was the lowest for **AC** and **S**. The results indicate that when there is little past information available for analysis, a small common use of an entity such as the name of a country makes it difficult to consider the text perfectly different owing to its influence.

## 4.5 Answers to the Fourth RQ: Can our algorithm find the main themes for texts?

The final experiment on the W2E dataset was the main theme analysis. The evaluation results indicated that our algorithm achieved 70% accuracy for allocating Wikipedia categories for the topics, whereas LDA achieved only 32% accuracy. Thus, LDA was inadequate for this analysis.

The queue and memory models achieved accuracies of 69.3% and 72.7%, respectively. Regarding the analysis time, the queue model took 4.82 seconds, whereas the memory model required only 1.44e-4 seconds. The results indicated that the memory model was better with regard to both the accuracy and the computing time.

## 4.6 Answers to the Fifth RQ: How well does our algorithm work for actual conversational texts?

Finally, we analyzed whether the proposed models produced appropriate results for actual discussion data. In summary, 1) our queue model judged the discussion to be appropriately transitional for 50% of the texts, and 2) the main theme analysis outputted reasonable results for 68% of the texts. Details are presented below.

In this experiment, we monitored several groups; thus, we used the queue model because of its short analysis time. First, we show what conversations actually occurred and what Wikipedia data was obtained. Note that the following discussion texts are English translations.

*The shape, or rather the shape, or rather the size of the inside of the tire, what is the speed, the speed per hour, the magnet, the magnet is different, it's a pool, Toyota. It's not that different.*

The Wikipedia articles obtained by applying ESA to above text were as follows. Our model used the Japanese versions of the Wikipedia articles; however, we list the corresponding English versions here.

- Magnet
- Mountains of magnets
- Nuclear magnetic resonance spectroscopy

We can see that all three articles retrieved by ESA were related to the above discussion text. After extracting the

**Table 5.** Percentage of category combinations with non-zero similarity

| | S | AA | BE | AC | LC | PE | IR | DA | HM |
|---|---|---|---|---|---|---|---|---|---|
| **S** | - | 0.26 | | 0.33 | 0.42 | 0.40 | 0.39 | 0.30 | **1.0** |
| **AA** | 0.67 | - | 0.33 | 0.66 | 0.5 | 0.42 | 0.54 | 0.46 | 0.66 |
| **BE** | 0.66 | 0.5 | - | 0.0 | 0.0 | 0.45 | 0.62 | 0.0 | 0.5 |
| **AC** | 0.75 | 0.25 | 0.0 | - | 0.0 | 0.53 | 0.42 | **1.0** | **1.0** |
| **LC** | 0.61 | 0.42 | 0.25 | 0.0 | - | 0.43 | 0.45 | 0.33 | 0.25 |
| **PE** | 0.54 | 0.48 | 0.12 | 0.33 | 0.47 | - | 0.54 | 0.46 | 0.43 |
| **IR** | 0.50 | 0.17 | 0.12 | 0.6 | 0.24 | 0.39 | - | 0.33 | 0.6 |
| **DA** | 0.4 | 0.38 | 0.25 | 0.0 | 0.33 | 0.39 | 0.37 | - | 0.0 |
| **HM** | 0.75 | 0.57 | 0.0 | 0.0 | 0.66 | 0.5 | 0.25 | 0.25 | - |

Wikipedia categories from these articles, our algorithm selected “magnet” as the main theme of this text.

The following text is a continuation of the above discussion.

*The engine thing is on, engine, how many speeds are you getting? meter, engine meter, engine meter, four minutes of time, working, finding each other, and from here on out, there’s not even this, this weird TV-like thing. It’s okay, though. Just a quick read. That’s about it. I’m sure that’s about it. There’s this thing that looks like a TV. There’s this thing that looks like a TV. There’s a battery-powered battery-powered thing. I don’t know. It’s called a Toyota.*

The following Wikipedia articles were obtained for the above text.

- Toyota City Library
- Toyota City
- Nyoi-ji Temple, Toyota City
- Radio-controlled car
- Dragon Nest
- Ufo Furusawa
- Yasuyuki Unezawa

Here, the first three articles are related to the word *Toyota* used at the end of the discussion text. The fourth article (“Radio-controlled car”) was also an appropriate result, as the discussion was about cars. Because the word *TV* was used in the text, some of the obtained articles were about actors and dancers working in the mass media, although these articles had nothing to do with the discussion. From the Wikipedia categories of the above articles, our model selected “Buildings in Toyota City” as the main theme.

As described above, when our model was applied to actual discussions, we can confirm that the obtained Wikipedia articles were relevant to the words in the text, and the Wikipedia categories selected as main themes were also relevant to the text.

The conversation text obtained from the actual discussion was divided into 1-minute segments, resulting in 20 conversation texts. We reviewed them to ensure that all the texts were appropriately transitioned in the discussion. Our model regarded appropriate transitions for 50% of the texts. We manually checked all the main themes selected by our model and found that the output was valid for 68% of the texts.

We analyzed the texts in which students engaged in appropriate discussions, but our models failed to recognize them as such. This analysis revealed that students frequently compared car components. For example, the first comparison was about the difference in tires between gasoline cars and electric vehicles, and soon the focus was on license plates and exhaust emissions. As the parts of the car are in close proximity to each other, the comparisons can be regarded as an exhaustive review of the objects of observation. However, the Wikipedia categories obtained from these texts were: “tires,” “chairs,” and “emissions trading.” As we clarified in Sections 1 and 3, our models assume that *the content of the discussion changes over time without gaps.* Under this assumption, our model analyzes the lexical consistency of the obtained Wikipedia categories. Thus, we consider that the above discussions violated the assumption of the models; there were no common Wikipedia articles or categories between the two texts, leading to failure of the analysis.

## 4.7 Discussions

*4.7.1 Model combination* The results of the discussion transition analysis revealed that both proposed models achieved higher accuracy in a shorter analysis time compared to the baseline. Reviewing the results of the main theme analysis revealed that the memory model outperformed the queue model for both the accuracy and the computing time. According to these results, the best option is to use only the memory model; however, in the W2E experiments, there was only one text to analyze at a time. In situations where multiple groups are conversing simultaneously, then analysis times may be longer than those reported in this paper, depending on the server performance. Thus, it may be suitable to use the queue model with a short analysis time for discussion transition analysis and the memory model for main theme analysis, or to use only the queue model. In the actual classroom experiments reported in this paper, we selected the queue model to reduce the server load.

*4.7.2 Written texts vs. colloquial texts* In this study, we used two types of data: the W2E dataset, which consisted of written text, including news articles published by news agencies, and actual conversational text, which was colloquial text, generated by elementary school students. We obtained appropriate Wikipedia data for both texts; however, there were differences in the analysis results. We obtained a high accuracy in the experiment using the W2E dataset, whereas the accuracy was lower for the actual conversational texts.

Because the W2E dataset included articles written by writers working for news agencies, many of the same entities tended to be used in sentences related to events grouped into a single topic. In contrast, the elementary school students' discussions tended to focus on the details after they had a comprehensive grasp of what they were learning, or they tended to focus on different points repeatedly. Our model was particularly effective for discussions where there was significant overlap between words used in previous or external information resources and new sentences because of our assumption. In addition, misspellings and omissions affected the accuracy for both the discussion transition and main theme analyses. The W2E dataset had no misspellings or omissions; thus, the Wikipedia articles we retrieved were adequate. In contrast, because we obtained the colloquial texts from the laptop and speech-to-text, there were missing words and other errors in the sentences. It is expected that future improvements in the accuracy of text-to-speech will enhance the effectiveness of the proposed models.

*4.7.3 Implications of the findings for teachers and schools* The findings obtained from the application of the proposed algorithms in actual classroom settings can be summarized in three key points. First, there is a need to address errors in the analysis results. Second, the use of the proposed algorithms leads to changes in teacher intervention. Third, the use of the proposed algorithms can assist in group formation.

The first point, as observed in Sections 4.6 and 4.7.2, is that the output of the proposed algorithms may contain errors. For instance, "Dragon Nest," the name of a video game, corresponds to an inappropriate Wikipedia article that does not accurately represent the discussion content. If one of the Wikipedia categories extracted from this article is selected as the main theme, the result may be incorrect. Teachers must be aware of this issue before using the proposed algorithms. Before the experiment, we informed the teacher that the system selects a category that is closest to the discussions from Wikipedia if a relevant category is missing. When the teacher used the system during an actual class, it displayed "magnets" as the main theme. This result puzzled the teacher because he did not expect that word to be related to the discussion content. When we asked the teacher for feedback after the class, he stated that he was initially surprised by the unexpected result. However, remembering the prior warning about potential inaccuracies, he focused more on checking if any warnings were displayed and whether the behavior of students seemed unusual, rather than concentrating solely on the main theme. Key findings from all observations indicate that teachers can conduct lessons more smoothly when they are aware of potential incorrect results. This includes not only errors from speech-to-text and ESA but also warnings generated by our algorithms. These warnings may arise in situations where discussions require an exploratory approach, emphasizing a holistic view—an insight drawn from the analysis of conversation transcripts of elementary school students. When a warning is issued, the likelihood of unnecessary interventions can be reduced by reviewing the actual conversation transcripts, utilizing existing methods described in Section 2.2 in combination, and limiting interventions to cases where warnings occur consecutively at least twice.

The second point, regarding the changes in teacher intervention due to the use of the proposed algorithms, demonstrates two significant effects. First, it allows teachers to monitor the status of multiple groups simultaneously, even without being physically present with the learners. Second, it helps prevent interruptions in discussions. The first effect is made possible by designing a user interface that displays the results of the proposed algorithms in a list format, enabling teachers to understand the discussion content in text form. The second effect occurs because teachers can observe the group dynamics without adding pressure by being physically close, thereby minimizing the need for intervention and promoting student autonomy. In the survey conducted with the teacher who cooperated in this experiment, it was noted that some students tend to stop speaking when the teacher approaches. The survey emphasized that the ability to monitor the status of the group, particularly the level of conversation activity is a valuable advantage.

The third point is that the proposed algorithms enable the teacher to gain early insights into the results of group discussions by visualizing the activity status of each group, even when information about the learners is insufficient. This provides valuable data that can be used for future group formations. In actual classroom group work, teachers typically consider strengths, weaknesses, personalities, and compatibility of all students when forming groups. However, at the beginning of a new academic year, there is little information about the students, making it difficult to form appropriate groups or anticipate which groups may need additional support. Even with a limited number of teachers, the proposed algorithms serve as an assistant for each group, allowing the teacher to listen closely to the conversations of students and providing alerts for appropriate intervention moments. This setup makes it easier for the teacher to monitor the status of all groups, even when prior information about the learners is lacking.

## 5 Conclusions & Future Work

We proposed models for teacher support to ease the simultaneous monitoring of several learning groups. We designed two objectives: 1) determine whether the discussion is appropriately transitioning and 2) identify the topics discussed. For the first objective, we proposed two models that assume that *discussion content changes over time without gaps*. The first model, called the queue model, limits the analysis to only the most recent discussion, and analyzes whether new discussions share similarities with the previous ones. The second model, called the memory model, reduces weights over time; however, it sums the weights per word at the end to assign higher weights to words that appear repeatedly. Using these models, we proposed algorithms for the second objective that allocate the main theme according to the Wikipedia categories closest to the center of gravity of the Wikipedia categories obtained for the queue model and those with the highest weights calculated by the memory model. To confirm the effectiveness of the proposed models for these two objectives, we measured the accuracy of the proposed models and baselines and found that the proposed models performed better in both cases.

In future studies, we plan to examine (a) *how our assumption that discussion content changes continuously is valid for several discussion types performed during classes.* We observed that elementary school students might react to the first thing they see rather than continuously shifting topics. Thus, we will investigate 1) the validity of this assumption for children and 2) its applicability to discussions between adults by analyzing a large corpus of discussion texts. In addition, we intend to propose (b) *models that capture situations in which many discussions are conducted that do not change continuously* by, for example, representing discussions using a tree structure. This model would be suitable for discussions that summarizes multiple topics that participants already discussed. Finally, we plan to extend (c) *the proposed model so that it does not classify input texts as deviations from the discussion when teachers or participants change topics while maintaining the main theme of the discussion or incorporate jokes and humor to enhance the motivation of other members.* This extended model can be implemented by integrating a classifier that determines whether an input text is an appropriate joke or humorous remark when the output of the discussion transition analysis in the current model exceeds certain thresholds. If the classifier determines that the text is appropriate, the output of the discussion transition analysis should be visualized in the same manner as when it remains within the threshold range.

### Acknowledgements

The authors would like to acknowledge everyone at elementary school for supporting this project and Google Japan G.K. for supporting this project through the "Research on developing a system for detecting the groups that need group work support."

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article. The authors have no relevant financial or non-financial interests to disclose.

## CRediT Author Statement

Yasunobu Sumikawa, Ryohei Ikejiri, and Yuhei Yamauchi contributed to the study conception and design. Algorithm was designed by Yasunobu Sumikawa and Kaoru Sekine. Implementation was performed by Yasunobu Sumikawa, Akikazu Takada, Akira Ichinose, Ari Murakami, and Yuki Toyono. Data collection and analysis were performed by Yasunobu Sumikawa, Ryohei Ikejiri and Meiko Sakasegawa. The first draft of the manuscript was written by Yasunobu Sumikawa and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

## Note

The sixth author has moved to Hiroshima University since completing the research.

## References

Ahmad M, Junus K and Santoso H (2022) Automatic content analysis of asynchronous discussion forum transcripts: A systematic literature review. *Education and Information Technologies* 27. DOI:10.1007/s10639-022-11065-w.

Ashikawa M, Kawamura T and Ohsuga A (2019) Proposal of grade training method for quality improvement in microtask crowdsourcing. *Web Intelligence* 17(4): 313–326. DOI:10.3233/WEB-190421. URL https://doi.org/10.3233/WEB-190421.

Atapattu T and Falkner K (2016) A framework for topic generation and labeling from mooc discussions. In: *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, L@S '16. New York, NY, USA: Association for Computing Machinery. ISBN 9781450337267, p. 201–204. DOI:10.1145/2876034.2893414. URL https://doi.org/10.1145/2876034.2893414.

Blei DM, Ng AY and Jordan MI (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.* 3(null): 993–1022.

Carbonell J and Goldstein J (1998) The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98. New York, NY, USA: Association for Computing Machinery. ISBN 1581130155, pp. 335–336. DOI:10.1145/290941.291025. URL https://doi.org/10.1145/290941.291025.

Chang MW, Ratinov L, Roth D and Srikumar V (2008) Importance of semantic representation: dataless classification. In: *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08. AAAI Press. ISBN 9781577353683, p. 830–835.

Chatterjee S and Dietz L (2021) Entity retrieval using fine-grained entity aspects. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380379, pp. 1662–1666. DOI:10.1145/3404835.3463035. URL https://doi.org/10.1145/3404835.3463035.

Chen B (2014) Visualizing semantic space of online discourse: the knowledge forum case. In: *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, LAK '14. New York, NY, USA: Association for Computing Machinery. ISBN 9781450326643, p. 271–272. DOI:10.1145/2567574.2567595. URL https://doi.org/10.1145/2567574.2567595.

Chu Z, Wang S, Xie J, Zhu T, Yan Y, Ye J, Zhong A, Hu X, Liang J, Yu PS and Wen Q (2025) LLM agents for education: Advances and applications. In: Christodoulopoulos C, Chakraborty T, Rose C and Peng V (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2025*. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7, pp. 13782–13810. DOI:10.18653/v1/2025.findings-emnlp.743. URL https://aclanthology.org/2025.findings-emnlp.743/.

Deerwester SC, Dumais ST, Landauer TK, Furnas GW and Harshman RA (1990) Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* 41(6): 391–407.

Ebbinghaus H (1987) *Memory : a contribution to experimental psychology*. Dover Publications.

Ezen-Can A, Boyer KE, Kellogg S and Booth S (2015) Unsupervised modeling for understanding mooc discussion forums: a learning analytics approach. In: *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, LAK '15. New York, NY, USA: Association for Computing Machinery. ISBN 9781450334174, p. 146–150. DOI:10.1145/2723576.2723589. URL https://doi.org/10.1145/2723576.2723589.

Ferragina P and Scaiella U (2010) Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10. New York, NY, USA: Association for Computing Machinery. ISBN 9781450300995, pp. 1625–1628. DOI:10.1145/1871437.1871689. URL https://doi.org/10.1145/1871437.1871689.

Hoang TA, Vo KD and Nejdl W (2018) W2e: A worldwide-event benchmark dataset for topic detection and tracking. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18. New York, NY, USA: Association for Computing Machinery. ISBN 9781450360142, p. 1847–1850. DOI:10.1145/3269206.3269309. URL https://doi.org/10.1145/3269206.3269309.

Hoffart J, Yosef MA, Bordino I, Fürstenau H, Pinkal M, Spaniol M, Taneva B, Thater S and Weikum G (2011) Robust disambiguation of named entities in text. In: Barzilay R and Johnson M (eds.) *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, pp. 782–792. URL https://aclanthology.org/D11-1072/.

Hsiao IH and Awasthi P (2015) Topic facet modeling: semantic visual analytics for online discussion forums. In: *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, LAK '15. New York, NY, USA: Association for Computing Machinery. ISBN 9781450334174, pp. 231–235. DOI:10.1145/2723576.2723613. URL https://doi.org/10.1145/2723576.2723613.

Le Q and Mikolov T (2014) Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14. JMLR.org, p. II–1188–II–1196.

Lieb A and Goel T (2024) Student interaction with newtbot: An llm-as-tutor chatbot for secondary physics education. In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703317. DOI:10.1145/3613905.3647957. URL https://doi.org/10.1145/3613905.3647957.

Liu B, Han FX, Niu D, Kong L, Lai K and Xu Y (2020) Story forest: Extracting events and telling stories from breaking news. *ACM Trans. Knowl. Discov. Data* 14(3). DOI:10.1145/3377939. URL https://doi.org/10.1145/3377939.

Machizawa D, Sawahata N, Ikejiri R and Sumikawa Y (2026) Storynetworks: An annotated dataset of event dependencies from short descriptions. In: *New Trends in Theory and Practice of Digital Libraries*. Cham: Springer Nature Switzerland. ISBN 978-3-032-06136-2, pp. 46–56.

Mendes PN, Jakob M, García-Silva A and Bizer C (2011) Dbpedia spotlight: shedding light on the web of documents. In: *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11. New York, NY, USA: Association for Computing Machinery. ISBN 9781450306218, p. 1–8. DOI:10.1145/2063518.2063519. URL https://doi.org/10.1145/2063518.2063519.

Ministry of Education, Culture, Sports, Science and Technology (2018) Koutou gakkou gakushuu sidou youryou kaisetsu chiri rekishi hen.

Mulla S and Shaikh NF (2024) Over comparative study of text summarization techniques based on graph neural networks. *Web Intelligence* 22(2): 231–248. DOI:10.3233/WEB-230014. URL https://journals.sagepub.com/doi/abs/10.3233/WEB-230014.

Nanni F, Ponzetto SP and Dietz L (2018) Entity-aspect linking: Providing fine-grained semantics of entities in context. In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18. New York, NY, USA: Association for Computing Machinery. ISBN 9781450351782, pp. 49–58. DOI:10.1145/3197026.3197047. URL https://doi.org/10.1145/3197026.3197047.

Prasojo RE, Kacimi M and Nutt W (2015) Entity and aspect extraction for organizing news comments. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15. New York, NY, USA: Association for Computing Machinery. ISBN 9781450337946, pp. 233–242. DOI:10.1145/2806416.2806576. URL https://doi.org/10.1145/2806416.2806576.

Qi Y, Zhou L, Si H, Wan J and Jin T (2017) An approach to news event detection and tracking based on stream of online news. In: *2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, volume 2. pp. 193–196. DOI:10.1109/IHMSC.2017.158.

Radinsky K and Davidovich S (2012) Learning to predict from textual data. *J. Artif. Int. Res.* 45(1): 641–684.

Raganato A, Camacho-Collados J and Navigli R (2017) Word sense disambiguation: A unified evaluation framework and empirical comparison. In: Lapata M, Blunsom P and Koller A (eds.) *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 99–110. URL https://aclanthology.org/E17-1010/.

Ramsdell J and Dietz L (2020) A large test collection for entity aspect linking. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368599, pp. 3109–3116. DOI:10.1145/3340531.3412875. URL https://doi.org/10.1145/3340531.3412875.

Sawahata N, Machizawa D, Ikejiri R and Sumikawa Y (2026) Llm-based dependency tracking for short event descriptions. In: *Linking Theory and Practice of Digital Libraries*. Cham: Springer Nature Switzerland. ISBN 978-3-032-05409-8, pp. 71–89.

Shen C, Wang D and Li T (2010) Topic aspect analysis for multi-document summarization. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10. New York, NY, USA: Association for Computing Machinery. ISBN 9781450300995, pp. 1545–1548. DOI:10.1145/1871437.1871668. URL https://

doi.org/10.1145/1871437.1871668.

Shochcho MI, Rahman MAU, Rohan S, Islam A, Heickal H, Rahman AM, Amin MA and Ali AA (2025) Improving user engagement and learning outcomes in llm-based python tutor: A study of pace. In: *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713958. DOI:10.1145/3706599.3720240. URL https://doi.org/10.1145/3706599.3720240.

Singh J, Nejdl W and Anand A (2016) History by diversity: Helping historians search news archives. In: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16. New York, NY, USA: Association for Computing Machinery. ISBN 9781450337519, pp. 183–192. DOI:10.1145/2854946.2854959. URL https://doi.org/10.1145/2854946.2854959.

Sumikawa Y, Takada A, Ichinose A, Murakami A, Toyono Y, Ikejiri R, Sakasegawa M, Sekine K and Yamauchi Y (2022) Online discussion transition analysis for group learning support. In: *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. pp. 249–255. DOI:10.1109/WI-IAT55865.2022.00043.

Tan Z, Zhang P, Tan J and Guo L (2014) A multi-layer event detection algorithm for detecting global and local hot events in social networks. *Procedia Computer Science* 29: 2080–2089. DOI:https://doi.org/10.1016/j.procs.2014.05.192. URL https://www.sciencedirect.com/science/article/pii/S187705091400369X. 2014 International Conference on Computational Science.

Wang T, Zhan Y, Lian J, Hu Z, Yuan NJ, Zhang Q, Xie X and Xiong H (2025a) Llm-powered multi-agent framework for goal-oriented learning in intelligent tutoring system. In: *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713316, pp. 510–519. DOI:10.1145/3701716.3715244. URL https://doi.org/10.1145/3701716.3715244.

Wang XJ, Lee CP and Mutlu B (2025b) Learnmate: Enhancing online education with llm-powered personalized learning plans and support. In: *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713958. DOI:10.1145/3706599.3719857. URL https://doi.org/10.1145/3706599.3719857.

Yang CC, Shi X and Wei CP (2009) Discovering event evolution graphs from news corpora. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 39(4): 850–863. DOI:10.1109/TSMCA.2009.2015885.

Yang Y, Urgo K, Arguello J and Capra R (2025) Search+chat: Integrating search and genai to support users with learning-oriented search tasks. In: *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400712906, pp. 57–70. DOI:10.1145/3698204.3716446. URL https://doi.org/10.1145/3698204.3716446.

Zarra T, Chiheb R, Faizi R and El Afia A (2018) Student interactions in online discussion forums: Visual analysis with lda topic models. In: *Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications*, LOPAL '18. New York, NY, USA: Association for Computing Machinery. ISBN 9781450353045. DOI:10.1145/3230905.3230920. URL https://doi.org/10.1145/3230905.3230920.

Zhang C, Lyu J and Xu K (2022a) A storytree-based model for inter-document causal relation extraction from news articles. *Knowl. Inf. Syst.* 65(2): 827–853. DOI:10.1007/s10115-022-01781-7. URL https://doi.org/10.1007/s10115-022-01781-7.

Zhang G, Zhu Z, Zhu S, Liang R and Sun G (2022b) Towards a better understanding of the role of visualization in online learning: A review. *Visual Informatics* DOI:https://doi.org/10.1016/j.visinf.2022.09.002. URL https://www.sciencedirect.com/science/article/pii/S2468502X22000924.

Zhou X, Gururajan R, Li Y, Venkataraman R, Tao X, Bargshady G, Barua PD and Kondalsamy-Chennakesavan S (2020) A survey on text classification and its applications. *Web Intelligence* 18(3): 205–216. DOI:10.3233/WEB-200442. URL https://journals.sagepub.com/doi/abs/10.3233/WEB-200442.