グループ学習での教師支援に向けた議論解析

澄川　靖信† 　高田　彰一†† 　一ノ瀬　晃††† 　村上　有弘†† 　豊野　勇紀††

池尻　良平†††† 　逆瀬川愛貴子†††† 　関根　　薫††††† 　山内　祐平††††

† 拓殖大学　〒 193–0985 東京都八王子市館町 815-1

†† Ddrive 株式会社　〒 799–0411 愛媛県四国中央市下柏町 426 番地 1

††† First Torrent 合同会社　〒 810–0801 福岡県福岡市博多区中洲 5-3-8 アクア博多 5 階

†††† 東京大学　〒 113–0033 東京都文京区本郷 7 丁目 3-1

††††† 東北大学　〒 980–8577 宮城県仙台市青葉区片平 2 丁目 1-1

E-mail: †ysumikaw@cs.takushoku-u.ac.jp, ††{takatter,ari,yuki}@ddrive.ai, †††ichinose@first-torrent.com,

††††{ikejiri,sakasegawa,yamauchi}@iii.u-tokyo.ac.jp, †††††kaoru.sekine@gmail.com

あらまし　グループ学習は高い学習効果があることが認められてる。特に近年では対面でのグループ学習だけでなく、オンラインでのコミュニケーションツールを用いて学習する機会が増加している。この変化によって、各グループの学習状況を教師が適切に把握できるように支援することの重要性も増している。本研究では、支援すべきグループを教師がすぐに把握できるように、グループごとの議論が適切に推移しているのか、一定時間ごとの議論主題を解析するアルゴリズムを提案する。これらの解析を実現するために、本手法は議論を記述したテキストから固有表現と Wikipedia カテゴリを抽出する。その後、解析対象を直近のものに限定するキューモデルと、繰り返される話題を重視するメモリモデルを適用する。また、Wikipedia カテゴリの中心的なもの、繰り返されるものを主題とみなす手法を適用する。本手法の有効性を評価するために、1 つの出来事に対する複数のニュースをまとめた W2E データセットと実際の会話データを用いたところ、高い精度で議論の解析ができることを確認した。

キーワード　議論解析, 遷移解析, topic detection and tracking, Wikipedia

1　はじめに

教育工学や学習科学の研究領域を中心に、グループワークはより高度な学習（high order learning）に効果的であることが示されている。これらの研究結果と連動するように、OECD や日本の新学習指導要領 [4] も、効果的なグループワークを授業中に行うことを強く求めるようになっている。

特に近年では対面でのグループ学習だけでなく、オンラインでのコミュニケーションツールを用いて学習する機会が増加している。対面授業とは異なり、オンラインツールはグループ活動を行うためにブレイクアウトルームといった他の参加者が会話に加わることが出来ない仕組みを利用する。このような仕組みを利用するとグループ活動を集中できるようになるものの、教師が一度に学習状況を把握できるグループ数は限られ、教師が全グループの活動を適切に把握することを難しくする。したがって、グループ学習や教師への支援を行うことの重要性が高まっている。

本研究では、グループごとの議論が適切に推移しているのか、どのような議題にしたがって議論しているのかを解析し、介入するのが望ましいと思われるグループを教師に伝えることを目的とする。

以下に本研究によって教師が介入するべき議論の様子の例を示す。

A:　鳥と恐竜の違いについて調べよう。
B:　俺は恐竜について調べるね。
A:　じゃあ鳥について調べる。
A:　生きているか絶滅しているかの違いはあるよね。
B:　確かに。あと、恐竜と鳥には骨格に違いはあるのかな。
A:　良い着眼点だね。今から骨格の違いについて調べよう。
B:　そういえば次の授業って何だっけ？
A:　さあ。
B:　あ、恐竜の特徴について最新の研究成果が報告されてる。
A:　その研究成果を読んでるけど、わからない。
B:　この成果はなんだろうね。

このグループは鳥と恐竜の違いについて議論している。骨格の違いに注目するまでは適切に議論している。しかし、各自で調べることにしてからは、議論とは関係が無い「次の授業って何だっけ」と B が発言している。その後、B が最新の研究成果を見つけたので、話の内容を元に戻しているが、A も B もその内容を理解することが難しく、議論を続けることが難しくなってしまった。上記の例の場合、「次の授業って何だっけ」と発言したような議論が飛躍した場合と、2 人とも調査内容がわからなくて議論が停滞した場合の 2 つの状況に対して、このグループが適切に議論できるように教師を呼ぶことが本研究の目的である。前者の場合、この例では紙面の都合上すぐに二人の会話は正常に戻ったことにしたが、もしこの状況が続くならば、学

習時間を浪費しないよう注意が必要である。

本手法は議論を記述したテキストから抽出した固有表現とWikipedia カテゴリが、これまでに解析した議題と共通部分を持つかどうか、それらの主題は何か、の 2 点を解析する。議論推移解析では、解析対象を直近のものに限定するキューモデルと、繰り返される話題を重視するメモリモデルを提案する。また、議論の主題を分析するために、取得したや Wikipedia カテゴリを座標空間に写像した後に重心を求める方法と、時間が経過しても繰り返されるものを主題とみなす方法を提案する。以上の手法を会話データが入力される度に実行し、介入が必要だと思われる確率を表す数値を出力する。この出力結果を教師が確認することで、複数のグループのうち、介入が必要だと思われるグループから優先的に対応することができる。

以上をまとめると、本研究は以下のリサーチクエスチョンを提案し、グループ学習中の教師を支援するためのアルゴリズムを提案する。

（1） 議論が適切に推移していることを解析できる?

（2） 関係のない話題の開始を検出できる?

（3） 停滞している議論を検出できる?

（4） 議論の主題を解析できる?

本手法の有効性を評価するために、1 つの出来事に対する複数のニュースをまとめたデータセットを用いた。議論の推移が適切に進んでいるかを解析したところ、98%もの高い精度を得られ、それらの主題を表す Wikipedia カテゴリも 70%という高い精度で取得できた。また、議論が逸脱・停滞しているデータセットを作成したところ、それらすべてを適切に検出できた。最後に、本手法を実際の小学校の授業で使用したときの結果を報告する。本稿のアルゴリズムは WI-IAT2022 [12] で発表したものと同一である。

WI-IAT2022 の原稿では大学院を修了した 3 名の日本人の議論に対して提案手法を実行した結果を報告したが、本稿では、最後に報告する小学校の授業での結果を報告する。

2 関 連 研 究

2.1 議 論 解 析

MOOCs などのインターネット上での学習環境が提案されて以来、Web 上での学習をより良いものにするための機械学習を用いた研究が行われている。議論内容を解析する研究としては、個々の学習者が発言したメッセージを構造化するための効果的な分類器の訓練方法を提案したもの [7] や、学習者同士のやり取りをネットワークとみなし、メッセージを送受信する学習者の理解度を推定する手法の提案 [5] が行われている。また、個々の学習者を解析対象とするだけでなく、グループを対象に、議論テキストに対してクラスタリングやキーワード付与を行って構造化した研究も行われている [1], [6]。

これらの先行研究では、本研究のような教師の支援を目指しているのではなく、学習者を支援することを目標としている。

グループ学習における教師支援の研究も行われている。Taoufiq *et al.* は学習者の議論の様子を可視化するために LDA によるトピックモデルを利用した可視化システムを提案した [14]。このシステムは各議論テキストに対して LDA を適用し、属する確率の高いトピックを構成する単語をその重みで大きさを変更して表示する。

このシステムも本研究と同様に教師支援のために議論テキストを解析するが、支援の目的が異なる。この先行研究はすべてのグループに対して適切なフィードバックを教師が行えるように支援することを目指している。一方、本研究は教師が介入するべきグループを発見することを目指している。この先行研究と本研究の目的は異なるので、同時に使用することが可能である。例えば、介入するべきグループが存在しない場合、先行研究のシステムを利用してより良い議論が行えるようにグループを支援することができる。しかし、途中で議論が上手くできなくなった場合は本研究を利用して速やかに教師がそのグループに参加し、問題の整理や学習者の発言を促す、という振る舞いを教師が行うことが可能である。

2.2 Topic Detection and Tracking

本研究は各グループが適切に議論を進めているかを解析する。これは、新しく入力された文章がこれまでの文章と同じ内容かどうかを分析することと同じである。このような分析は、機械学習の研究では topic detection and tracking (TDT) として研究されてきた。Radinsky と Davidovich は将来起こり得る出来事を予測するために、過去に生じた出来事を報じるニュースから因果関係を予測する TDT の手法を提案した [11]。この因果関係を発見するとき、新しく入力されたニュースが、これまでに報道されたニュースと同じ内容であるかどうかを分析する。もし同じだと判断した場合、これまでのニュースとして蓄積した鎖にこの新しいニュースを関連付ける。この手順は本研究の議論の推移分析と同じである。もし新しく入力された議論テキストがこれまでの内容と同じであれば、その新しいテキストを蓄積する。さもなければ、教師を呼びだすための手続きを行い、このテキストを破棄する。TDT の先行研究では、オンラインニュースから話題となっているニュースの検出と追跡を行う手法 [10]、コミュニティ検出を利用した世界的な話題になっているニュースや特定の地域のみで話題になっているニュースの検出を行う手法 [13] が既に提案されている。上記の先行研究は出来事の予測を目指しているので、議論推移分析にそのまま適用することは難しい。本研究は教師支援を目指しているので、これまでのテキストと内容が異なるテキストが入力された場合には教師にその結果を知らせることが重要である。

3 議論推移解析アルゴリズム

図 1 に本手法の全体像を示す。提案手法を適用する前に、議論推移を解析する学習者の会話データは既にテキストに変換されていることを仮定する。本手法は、まず、入力された会話テキストに対して、固有表現と Wikipedia カテゴリの 2 種類を抽出する。これらを用いる議論推移分析は、入力された文章が適切に議論が推移している、これまでの内容から飛躍している、

これまでの内容と同じ停滞である、の 3 種類を分析する。もし、入力されたテキストが適切に議論を行えていると解析できれば、以降の解析でこの結果を利用できるように記録する。その後、教師が介入できるように、リスク確率を表す数値を返戻する。このとき、グループごとの議論内容を容易に確認できるように、各テキストの主題を分析する。本手法の解析結果である 2 つの確率の値と主題は、教師が確認する画面に提示する。

以降では、まず、議論テキストから固有表現や Wikipedia カテゴリを抽出する解析方法について述べる。その後、この解析結果を利用する議論推移解析と主題解析の実現方法について述べる。

3.1 固有表現& Wikipedia カテゴリ解析

本研究はテキストの中に実際に含まれている単語と、そのテキストの意味を表す単語の 2 種類を用いる。前者のものとして、固有表現を利用する。本手法は固有表現の検出は既存手法を利用することを仮定する [1]。後者のテキストの意味を表す単語として、本手法は Wikipedia のカテゴリを利用する。テキストから Wikipedia カテゴリを取得するために、本研究では explicit semantic analysis (ESA) [2] を利用する。ESA は文章全体に対して意味的に近い Wikipedia 記事を検索し、それらをランキング形式で出力する。ESA によって会話テキストと関係性が高い Wikipedia 記事を取得できる。Wikipedia 記事には Wikipedia 編集者によって定義されているカテゴリが付与されている。本研究ではこれらのカテゴリの名称はその記事を抽象化しているものとみなす。すなわち、この Wikipedia カテゴリの集合は入力された会話テキストを抽象化したものとみなす。しかしながら、Wikipedia カテゴリはその記事に関係するなら付与されるので、議題とは関係がないカテゴリも存在する。このようなノイズを取り除くために、本手法では、議論のテキストから抽出した名詞を含むものだけを解析対象のカテゴリとする。

この処理によっ固有表現と Wikipedia カテゴリを得られたので、以降の解析ではこれらを単語リスト *words* に含めて利用する。

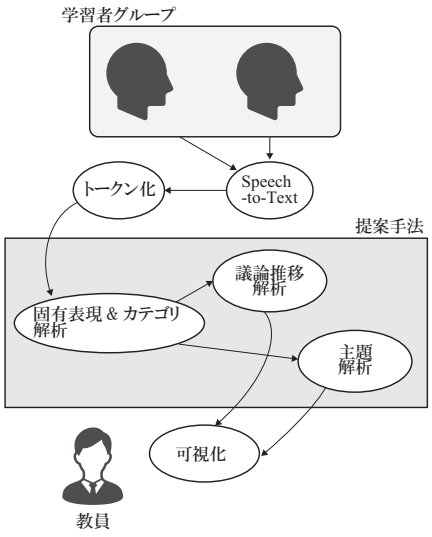

図 1　システム全体像

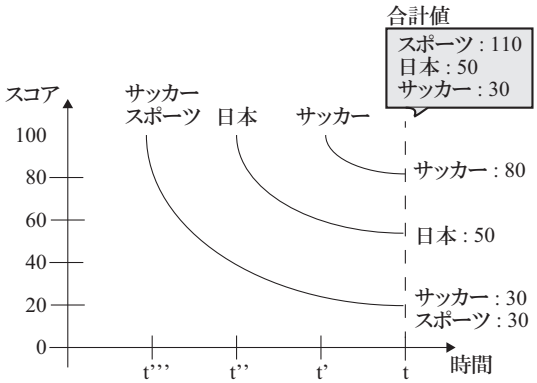

図 2　忘却曲線モデル

3.2 議論推移解析

本研究では議論推移を分析するためにキューモデルとメモリモデルの 2 つを提案する。2 つのモデルは共に「話題は時間の経過と共に変化する」と仮定している。この変化を捉える方法が 2 つのモデルで異なる。キューモデルは分析対象から古いテキストを除外するモデルである。メモリモデルは「話題は時間の経過と共に変化するが、繰り返し出現する話題は議論内容の中心的なものである」の仮定の下、新しい解析結果に重みを付与するが、それらの値は時間が過ぎると重みを減らすモデルである。これらの 2 つのモデルの詳細を以降で記述する。

3.2.1 キューモデル

キューは先入れ先出しの規則に従って、新しいオブジェクトが入力されたら、一番古いオブジェクトを除外するデータ構造である。本研究では、新しいテキストが出現するたびに古いテキストを除外して、解析対象を直近 N 個に限定することで最新の議論を分析する。このモデルでは一様分布を仮定し、分析対象の N 個のテキストに関しては平等に解析する。すなわち、新しいテキストと N 個前の分析結果に共通するものがあるかどうかで議論推移を分析する。もし共通する結果が一定以下であれば、議論の推移が飛躍したとみなす。一方、共通する結果が一定以上あれば、議論が停滞しているとみなす。

式 1 にキューモデルによる議論推移分析を示す。

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

$Words(N)$ は過去 N 回分の議論テキストの単語を取得する関数である。$\delta(w, Words(N))$ は、過去 N 回分の議論テキストに w が含まれていれば 1 を返し、さもなければ、0 を返す関数である。この式の分子が過去のテキストと入力された新しいテキストが共通して持つ単語の数を数える。分母は解析対象のすべてのテキストに含まれる単語の数を表すので、この割り算によって数値を確率として表現している。キューモデルでは、この式の値が指定の範囲内であれば、新しいテキストは議論テキストに追加され、1 番古いテキストを除外する。

1：本研究の実装では AutoML を使用した。https://cloud.google.com/natural-language/automl/docs

3.2.2 メモリモデル

このモデルでは、長期間に渡り繰り返し議論されている内容は、その時間で特に重要なテーマであると仮定する。この仮定を実現するために、繰り返し学習によって内容が保持される様子を表す忘却曲線 [3] によって本モデルを表現する。

本モデルでは、図 2 に示すように、*words* に含まれる単語に対して重みを計算する。この図は時刻 `t'`, `t''`, `t'''` で 3 つのテキストが生成され、それらのテキスト中に “soccer”, “sports”, “Japan” の 3 種類の単語が含まれていて、それらの重みの計算方法を示す。このモデルは、各単語の重みを、時間の経過とともに指数関数的に低減させている。しかし、繰り返し出現する単語を高く出来るように、単語の出現ごとに重みを求め、最後にそれらを総和する。

以降では上述したアルゴリズムを定式化する。この重みを計算するために、忘却曲線をモデル化した指数関数 exp を使用する。記憶は時間が経つにつれて覚えている確率が下がる傾向があるので、本モデルでも、現在の時刻 t と前回のテキストが現れた時刻 t' の差が大きいほど値は小さくする。この現象を表現するために、2 つの時間の差を exp の引数として与える。新しいテキストが入力される度に、既に解析したテキストは以前よりも時間が経過しているので、それらの重みを再計算する。また、テキストごとに同じ単語でも異なる重みとするので、それらの総和を、t におけるその単語の最終的な重みとする。

以下にメモリモデルでの重み低減を実現する数式を示す。

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

過去のテキスト 1 つ 1 つを解析して、現在、最も重要である話題を見つけるために、次式に示すように、式 2 の総和を求める。

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

w は単語、関数 $PText(t)$ は時刻 t よりも前の議論テキストを記録した時刻を取得する、関数 δ は w が $text(t')$ に含まれるなら 1、含まれないなら 0 を返す。

最後に、すべての過去のテキストを解析した結果をまとめる。この最終結果は次式によって計算できる。

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

キューモデルと同様に、このスコアが 2 つの閾値との比較で結果を決める。この値が適切に議論が推移していることを示す場合、時刻 t でのテキストとして解析できるように、$PText(t)$ にテキストを記録する。

3.3 主題解析

各テキストの主題を解析するために、ESA で取得した Wikipedia カテゴリの中から適切なものを選ぶ。Wikipedia カテゴリは、本来は Wikipedia 記事を構造化するために導入されたが、カテゴリはその記事を端的に表現する特徴を持つ。この特徴を利用して、本手法は議論のテキストに対して得られた Wikipedia カテゴリはそのテキストを抽象的に表しているとみなし、その中から主題となるものを選択する。この選択方法は議論推移モデルごとに異なる手順で各カテゴリを数値化し、適切なものを主題として 1 つ選択する。

Algorithm 1 擬似コード

Input: 議論テキスト *txt*, 過去の議論テキスト *N*, 閾値 *thre*
Output: スコア *score*, 主題 *m*

```
1:  Function DiscussionAnalysis(txt, N)
2:    // テキスト解析
3:    entities ← ExtractEntities(txt)
4:    WikiCats ← ESA(txt)
5:    words ← entities + WikiCats
6:    // 議論推移解析
7:    score ← 式 1 または式 4 を適用した結果
8:    if score > thre
9:      N ← txt // txt を過去の議論テキストとして保存
10:   end if
11:   // 主題解析
12:   m ← 節 3.3 の手法の適用結果
13:   return score, m
```

3.3.1 キューモデル

キューモデルは一定の範囲の直近のテキストを全て解析対象とするので、全てのカテゴリを座標空間に埋め込む。ここで、Wikipedia カテゴリは 1 つ以上の Wikipedia 記事に対して付与されているので、それらの記事がそのカテゴリを詳細に表す文章とみなす。本手法は、全ての Wikipedia 記事を用いて Doc2Vec [9] を訓練する。その後、各カテゴリの全記事を 1 つの文章とみなして、Doc2Vec モデルを適用して座標空間に埋め込む。この結果、取得したカテゴリに対してベクトルを求めることができるので、それらの重心を計算することが可能になる。取得したカテゴリの中からこの重心に最も近いものを主題とみなす。

3.3.2 メモリモデル

メモリモデルは式 4 によって、各 Wikipedia カテゴリに対してスコアを計算する。長い期間に渡って繰り返し出現する Wikipedia カテゴリはこのスコアが高くなるので、このモデルの場合は式 4 の値が最も高いカテゴリを主題とする。

3.4 アルゴリズムの全体像

Algorithm 1 に本手法の疑似コードを示す。本手法は、まず、入力として与えられたテキスト *txt* に対して、テキスト処理として固有表現と Wikipedia カテゴリの抽出を行う（2~4 行目）。5 行目でこれらの結果を統合したものに対して、Sec. 3.2 で述べたキューモデル、メモリモデルのいずれか片方を適用して、適切に議論が推移している確率を求める（6~7 行目）。もしこの確率が閾値を超えていれば、適切に議論が行われているとみなし、この文章を議論を構成するテキストとして記録する（8~10 行目）。最後に、11~12 行目に示すように、入力されたテキストの主題を解析して、その結果と議論推移の確率値を返戻する。

4 実　　験

4.1 リサーチクエスチョン

本稿では、提案手法の有効性を評価するために、以下のリサーチクエスチョンに従って実験する。

RQ1　適切に推移している議論を分析できた？
RQ2　停滞している議論を検出できた？
RQ3　急に出現した関係がない話を検出できた？
RQ4　主題を表す Wikipedia カテゴリを適切に取得できた?
RQ5　実際の会話テキストに対しても適切に推移解析できた？

4.2 設　　定

4.2.1 RQ1 と RQ4 のデータセット

RQ1 と **RQ4** は TDT 問題とみなすことができるので、TDT のデータセットとして作られた W2E dataset [8] を用いて定量評価を行う。

W2E はロイター通信、New York Times、BCC などの各新聞報道社が報じたニュースの中から、同一の出来事に対するものを時系列に並べてトピックとしてまとめている。このデータセットに含まれるトピックは 2016 年に報道されたものに限定しているが、3,083 トピックも利用できる。W2E データセットは複数の出来事のリストをトピックとしてまとめている。これらの出来事は英語版 Wikipedia の Current events ポータルに記録されたものである。W2E データセットの出来事は、アメリカ大統領選挙、英国の EU 離脱、夏季オリンピック、といったものを含んでおり、それらの出来事に対して、Wikipedia 編集者が手動で定義した 10 個の出来事カテゴリに属している。W2E データセット構築は手動で上記の出来事を手動でトピックとして構成できるものを集約し、出来事カテゴリをトピックに対して付与している。

W2E データセットは TDT のベンチマークデータセットとして作成されたが、1 つのトピックに 1 つの出来事しか含まれていないものが存在する。本手法の有効性を適切に評価するために、少なくとも 3 つの出来事が含まれるトピックだけを本評価で使用する。この選別を行ったところ、1,781 個の出来事を含む、269 個のトピックを抽出できた。また、本稿では、上記の 10 カテゴリのうち、以下の 9 カテゴリだけを抽出できた。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**).

本稿で抽出した W2E データセットの統計情報を表 1 に示す。表 2 はカテゴリごとの出来事数、トピック数、文章中のトークン数の平均値を示す。また、これらのリサーチクエスチョンのために本稿で使用した W2E データセットのサブセットは誰でも再検証できるように公開している [2]。

表 1　テストデータセットの統計情報.

トピック数	269
出来事数	1,781
トークン数の平均	32.9
トピック毎の出来事数の平均	6.62
RQ5 のテーマ数	7
RQ5 の文字数	7,081

表 2　カテゴリ毎のテストデータセットの統計情報

	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

4.2.2 RQ2 のデータセット

RQ2 の評価のために、あるトピックに含まれる 1 つの出来事を意図的に複製し、それらを続けて並べることによって停滞している議論を再現できるように W2E データセットを拡張した [3]。この評価では **RQ1** と同様に 269 トピックを使用する。

4.2.3 RQ3 のデータセット

RQ3 の評価では、2 つのトピックに対して、1 つのトピックをもう一方のトピックの最後に挿入して新しい 1 つのトピックを定義することで、急に出現した関係がない話を再現できるように W2E データセットを拡張した [4]。この操作をすべての 2 つのトピックの組み合わせに対して行った。しかし、全く関係が無い 2 つのトピックを混ぜることを保証するために、次の 2 つの条件のうち片方でも成立する場合は混ぜることを中止した。

- 追加する出来事の文章と追加されるトピックの出来事の文章の両方に共通の単語が存在する。
- 追加する出来事の文章に対して作成した LSA による特徴ベクトルが、追加されるトピックの各出来事の文章に対して作成した LSA による特徴ベクトルに依存している。

2 つ目の条件を検査するために、本稿では次式で示す相互情報量を利用した。

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

もし相互情報量の値が 0.3 よりも大きい値ならば、依存していると判断した。以上の手続きを行ったところ、**RQ3** の評価のために、2,055 トピックを作成した。

4.2.4 RQ5 のデータセット

最後に、実際に議論している会話データに対しても提案手法が適切な結果を得られるかを分析する。この分析は、小学校の社会科の授業として日本語話者の小学生がガソリン車と

2：https://drive.google.com/file/d/1ob5QvQxavwTSAcw720QR6SwLFU0Af0ZR/view?usp=sharing

3：この拡張データセットは次の URL で公開している。https://drive.google.com/file/d/1ylEkYfVwzKCwfkwDQJ6qqEW7c2ojxns-/view?usp=sharing

4：この拡張データセットは次の URL で公開している。https://drive.google.com/file/d/1IItlpa0kebEtiC6Do7_KDc0KsJUGAzjn/view?usp=sharing

表 3　**RQ2.** の結果

	TDT	キュー	メモリ
Correct ratio	28.5%	96.6%	97.0%
Transition analyzing time (sec.)	5.52e-06	0.08	1.59

表 4　**RQ2.** でのエラー分布

	AA	**PE**	**IR**	**S**	**LC**	**DA**	**BE**
キュー	2.8%	4.2%	2.7%	2.2%	5.2%	3.7%	4.5%
メモリ	2.2%	3.9%	2.7%	2.2%	4.3%	3.7%	4.5%

EV 車の違いについて議論した様子を、google cloud platform の speech-to-text によってテキスト化した文章を用いる。この授業は、10 分間の各グループでの議論、5 分間の教師による全グループの現状確認、5 分間の各グループでの議論、の計 20 分で実施した。グループでの活動では、配布された資料を参考にしながら 2 人一グループで議論した。各グループには 1 台の端末を机においてもらい、端末から音声データを取得し、これを実際の会話データとしてサーバーに保存した。取得した会話データは 1 分単位で区切ることで、教師が 1 分ごとの議論の様子を確認できるようにした。

4.2.5　パラメータ

各パラメータ調整のために、著者らが実際に議論したテキストを用いた。キューモデルの N の値を 4 にした。α と β はそれぞれ 0.02、0.3 とした。

4.2.6　比 較 対 象

本稿では以下の手法を比較対象として用いる。

LDA:　LDA のみを使って主題を分析する。

TDT:　Kira radinsky らが提案した TDT 手法 [11] である。この手法はコサイン類似度を用いて集めた互いに類似するテキストに対して、固有表現によるエントロピーが高まる文章を選ぶ。

4.2.7　評 価 基 準

RQ1 は W2E のトピックを構成する出来事に対して、適切に推移していると提案手法が判断した数で評価した。**RQ2**~**RQ3** は意図的に挿入したノイズを適切に検出できた数で評価した。**RQ4** は著者らが全ての結果を手作業で確認した。**RQ5** は上記の **RQ1**~**RQ4** と同じ手続きで結果を求めたが、その確認作業は著者らと 3 名の協力者による手作業で行った。

4.3　W2E データセットでの結果

RQ1. 議論推移分析は適切に推移している議論を分析できた？
A1. キューモデルもメモリモデルも約 97%のテキストを正しく解析できた。
A2. メモリモデルの方がキューモデルよりも多くのテキストを正しく解析できた。

表 3 に提案手法のモデルが適切に推移している議論での解析結果を示す。まず、比較対象と比べると提案手法は高い精度が得られたことがわかる。2 つの提案手法は共に正解率が約 97%という高い結果を得た。適切に推移している議論に対して、不適切な推移をしたと間違って判断した数に注目すると、キューモデルが 50 だったのに対してメモリモデルは 44 だったので、メモリモデルの方が高い精度が得られたことがわかる。一方、議論推移の分析に要した解析時間は、キューモデルが 0.1 秒未満だったが、メモリモデルは約 1.6 秒だった。キューモデルは解析対象を直近のものに限定して出現数を数える一方、メモリモデルは過去の議論テキストで出現したものに対して指数関数で重みを計算するので、計算コストが重い傾向がある。しかし、解析時間が約 1 秒と短い時間で完了したので、実践的には大きな問題ではない。

表 4 に提案手法の 2 つのモデルが、どのようなカテゴリの出来事に対して間違った判断をしたのかを分析した結果を示す。2 つのモデルは共に各カテゴリにおいて約 2~5%の出来事で誤解析をしていた。キューモデルは **LC** の出来事に対して、メモリモデルは **BE** の出来事に対して、それぞれ最も誤解析率が高かった。

RQ2. 議論推移分析は停滞している議論を検出できた？
A. 99.6%の文章に対して適切に停滞を検出できた。検出に失敗した理由は、有効な固有表現と Wikipedia 記事の取得に失敗していたことが考えられる。

この分析では、メモリモデルとキューモデルの両方が共に、意図的に挿入した冗長な出来事 269 個の中から 268 個（99.6%）に対して、適切に停滞していると検出した。以降では、停滞を検出出来なかった理由について分析する。まず、268 個の適切に停滞を分析できた出来事の文章から抽出できた名詞、Wikipedia 記事、Wikipedia カテゴリ、固有表現について調べたところ、それぞれの平均値は、10.9、24.9、170.1、5.4 であった。一方、この 4 つの項目に関して、停滞を検出出来なかった出来事の文章から抽出できた数は、それぞれ、6.0、10.0、74.0、0.0 であった。すべての数値が正解できた出来事のものより少ないことがわかる。特に抽出できた固有表現が 0 個だった。提案手法が使う単語数が少ないと解析に失敗することがわかる。

次に、この出来事を含むトピックに対する提案手法の解析結果を確認した。解析に失敗した出来事の文章は “The Syrian cessation of hostilities truce is in effect, as of midnight, Saturday, local Syrian time” であった。また、このトピックには 4 つの出来事が含まれており、いずれもシリアで生じた出来事であった。上記の出来事は重複させているので、このトピックには 4 種類の 5 つの出来事を含む。最初の 3 つの出来事に対して抽出した Wikipedia カテゴリと固有表現に注目したところ、上記の出来事と共通するものは存在しなかった。そのため、4 つ目の出来事に対して本手法は「急に出現した関係がない話」と誤った判断を下していた。この結果から、4 つ目の出来事に対して得られた Wikipedia カテゴリと固有表現を保存せず、もう一度同じテキストを解析し、もう一度、「急に出現した関係がない話」と間違った結果を下したことがわかる。

図 3　誤解析したカテゴリの組合せ

RQ3. 議論推移分析は急に出現した関係がない話を検出できた？
A. すべての文章に対して、正しく、議論の推移が不適切であると検出できた。

この分析結果もメモリモデルとキューモデルの両方が同じ結果であった。2,055 個の分析対象である出来事のすべてに対して、正しく、議論の推移が不適切であると判断できた。この結果を詳細に分析すると、2,055 個中、完璧に議論が異なると解析できたものは 1,150 個だった。残りの 905 個の出来事は、これまでの推移とはあまり関係が無いと判断した。

図 3 にこれまでの推移とはあまり関係が無いと判断した 905 個の出来事を含むトピックに対して、どのカテゴリのトピックから選んだ出来事を混ぜたのかを示す。割合が特に高いカテゴリは、**S** に **HM** のものを混ぜた場合、**AC** に **DA** や **HM** を混ぜた場合であって、それらはほぼすべてが該当した。表 2 に注目すると、**S** と **AC** は 1 つのトピックに含まれる出来事数の平均値が約 3~4 個であり、9 個のカテゴリの中で平均値が最も低い 2 つのカテゴリだった。この値が低いと、議論推移状況を分析するために利用できる情報が少なくなるので、少しでも共通する固有表現や Wikipedia カテゴリが存在すると、その影響が大きくなる。これが「完璧に議論が異なると解析」ではなく、「これまでの推移とはあまり関係が無い」と判断した理由と考えられる。

RQ4. 主題を表す Wikipedia カテゴリを適切に取得できた？
A. 提案手法は 70%の精度で適切なカテゴリを取得できた。

W2E データセットの中からランダムに 100 個の出来事を選び、それらに対して各モデルを適用した結果を評価した。本研究はグループ学習中の教師への支援を目的としているので、モデルが付与したカテゴリ名と出来事文章に共通部分があるかどうかを評価基準とした。もし出来事の記述文章中に同じ固有表現があれば正しいとみなす。また、固有表現が土地名（例：東京）だが Wikipedia カテゴリがその国名（例：日本）であっても正しいとした。

比較対象である LDA のトピックから主題を割り付けるアルゴリズムの正解率は 32%であった。提案手法であるキューモデルとメモリモデルの正解率は、それぞれ、69.3%と 72.7%であった。従って、主題判定を行うためには LDA を適用するだけでは難しいが、本手法は多くの場合に良い結果を得られていることがわかる。次に、2 つの提案手法の解析時間を調べると、キューモデルは 4.82 秒、メモリモデルは 1.44e-4 秒だった。キューモデルは座標空間上に写像した Wikipedia カテゴリの特徴ベクトルに対して、それらの重心を求める計算を行うため、解析結果の上位を取り出すキューモデルよりも多くの計算時間が必要である。しかし、キューモデルも 5 秒以内に結果が出ているので、実際の使用状況を考慮しても問題にならない程の早さで結果が得られている。

4.4　実際の議論テキストでの結果

RQ5. 議論推移分析は実際の会話テキストに対しても適切な結果を得られた？
A1. 50%のテキストに対して適切に議論が推移していると判断し、誤って飛躍や停滞と判断したものは無かった。
A2. 主題付与は 68%に対して妥当な結果を出力した。
A3. 教師に誤った結果を提示しても、机間巡視に違いが起きた。

最後に実際の会話データに対して提案手法を適用した結果を分析する。以下に、実際に議論中に発言された文章を示す。

> 形っていうか 形 っていうかなんか タイヤの中の大きさっていうなん 何 時速 時速 磁石 磁石が違う プールだとトヨタ。 そんなに変わらなくない。

この文章に対して ESA を適用して得られた Wikipedia 記事は以下の通りである。

- 磁石
- 磁石の山
- 核磁気共鳴分光法

「核磁気共鳴分光法」は磁石に関係しているので、取得できた 3 つの記事は全て上記の会話テキストに関係していた。また、これらの記事から Wikipedia カテゴリを抽出した後、この文章の主題として「磁石」を本手法は選択した。

上記の会話の続きを以下に示す。

> エンジンのやつ付いてるよ エンジン 何速 出てる マスオ」ってメーター エンジンのメーターでエンジンのメーター 4 分間時間 働いてね 見つけ合いながら ここから これもないよ変なテレビみたいなこの。 大丈夫ですが。 ちょっと 一読 このぐらいだろう。 実装。 テレビみたいなやつがね テレビみたいなやつがある バッテリー式バッテリー式のやつがある。 よくわからん 豊田っていう

この文章に対して次の Wikipedia 記事を取得した。

- 豊田市中央図書館

- 豊田市
- 如意寺 (豊田市)
- ラジコン模型自動車
- ドラゴンネスト（ゲーム）
- 古澤侑峯
- 釆澤靖起

これらの記事のうち、最初の 3 つは会話テキストの最後の方にある「豊田」と関係している。また、この会話内容は車に関するものなので 4 つ目の記事の「ラジコン模型自動車」も適切な結果である。テキスト中に「テレビ」という単語があるので、マスメディアで活躍する俳優や舞踊家といった議論内容とは関係がない記事も取得していた。以上の記事から「豊田市の建築物」を主題として選択した。

以上のように、実際の議論で取得したテキストに対して本手法を適用すると、テキスト中にある単語から関連性がある Wikipedia 記事を取得できていること、主題として選択する Wikipedia カテゴリも一定程度の関連性があることが確認できる。

実際の議論で取得した会話テキストを 1 分ごとに分割したところ、20 回分の会話テキストが得られた。これらすべての結果を確認したところ、すべてのテキストが適切に議論が推移していることが確認できた。本手法を適用したところ、50%のテキストに対して適切に議論が推移していると判断し、誤って飛躍や停滞と判断したものは無かった。各テキストに付与した主題を確認したところ、68%に対して妥当な結果を出力していた。

議論は適切に推移しているが提案手法が解析に失敗した内容について分析したところ、学習者は車の部品について比較していることが多かった。例えば、最初はガソリン車と EV 車のタイヤの違いについて比較し、すぐに着眼点がナンバープレートや排気ガスに着目していた。これらは車の違いに近い場所にあるため、議論を開始したときの網羅的に観察対象を確認している様子とみなすことができる。しかし、これらの文章から Wikipedia のカテゴリを取得したところ、「タイヤ」、「椅子」、「排出取引」が得られた。本手法は得られた Wikipedia カテゴリの字面の一致性を解析するため、これらの共通点である車が存在しなかったので、誤って議論が飛躍したと解析した。

実際に授業を運営していた教師に本システムが出力した結果に対する印象を確認したところ、「磁石」のような授業内容と異なる結果が主題として表示されると戸惑うことがわかった。しかし、事前に Wikipedia カテゴリとして近そうなものが選ばれることを伝えていたので、多少のずれや間違いを気にするのではなく、一度に複数の児童達の様子を確認できる点に注目し、様子を確認するグループの順番が本システムの有無で違いがあった。

5 まとめと今後の課題

本稿ではグループ学習としての議論を行うとき、議論が適切に行えていないグループを教師が容易に発見できるように支援するためのアルゴリズムを提案した。本手法は、「時間の経過と共に議論内容が変化する」という仮定の下、各グループの議論のテキストが発生する度にこれまでの内容と同じ内容かどうかを分析する 2 つのモデルを提案した。また、各議論のテキストがどのような内容なのかを表す Wikipedia カテゴリを付与するアルゴリズムを実現した。上記の有効性を評価したところ、キューモデルは短い時間で議論推移分析の結果を得られるが、メモリモデルの方が議論推移分析の精度が高いことが明らかになった。議論テキストが表す内容を付与するアルゴリズムは、メモリモデルで得られた数値を利用する場合は直ちに結果を得られるが、重心を求めるアルゴリズムの方が精度が高いことが明らかになった。

今後の課題として、議論の質評価が考えられる。現在のアルゴリズムは、議論が連続的に変化することを仮定しているが、実際の議論では、既に出た複数の議題をまとめる質の高い議論が行われることがある。議論内容を線形的に解析するのではなく、木構造として表現することで、議論の推移を評価することが考えられる。

謝辞：本研究は Google 社の助成を受けたものである．

文　献

[1] Atapattu, T., Falkner, K.: A framework for topic generation and labeling from mooc discussions. pp. 201–204. L@S '16 (2016)

[2] Chang, M.W., Ratinov, L., Roth, D., Srikumar, V.: Importance of semantic representation: Dataless classification. pp. 830–835. AAAI'08 (2008)

[3] Ebbinghaus, H.: Memory : a contribution to experimental psychology. Dover Publications (1987)

[4] Ministry of Education Culture, S.S., Technology: Koutou gakkou gakushuu sidou youryou kaisetsu chiri rekishi hen (2018). URL http://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2018/08/29/1407073_03_1.pdf

[5] Erlin, Susandri, Yenni, H.: Social network analysis for online discussion: Number of links vs. sum of weight. pp. 82–86. ICCIS '16 (2016)

[6] Ezen-Can, A., Boyer, K.E., Kellogg, S., Booth, S.: Unsupervised modeling for understanding mooc discussion forums: A learning analytics approach. pp. 146–150. LAK '15 (2015)

[7] Ferreira, M., Rolim, V., Mello, R.F., Lins, R.D., Chen, G., Gašević, D.: Towards automatic content analysis of social presence in transcripts of online discussions. pp. 141–150. LAK '20 (2020)

[8] Hoang, T.A., Vo, K.D., Nejdl, W.: W2e: A worldwide-event benchmark dataset for topic detection and tracking. pp. 1847–1850. CIKM '18 (2018)

[9] Le, Q., Mikolov, T.: Distributed representations of sentences and documents. pp. II–1188–II–1196. ICML'14 (2014)

[10] Qi, Y., Zhou, L., Si, H., Wan, J., Jin, T.: An approach to news event detection and tracking based on stream of online news. pp. 193–196 (2017)

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

[12] Sumikawa, Y., Takada, A., Ichinose, A., Murakami, A., Toyono, Y., Ikejiri, R., Sakasegawa, M., Sekine, K., Yamauchi, Y.: Online discussion transition analysis for group learning support. WI-IAT'22 (accepted)

[13] Tan, Z., Zhang, P., Tan, J., Guo, L.: A multi-layer event detection algorithm for detecting global and local hot events in social networks. Procedia Computer Science **29**, 2080–2089 (2014)

[14] Zarra, T., Chiheb, R., Faizi, R., El Afia, A.: Student interactions in online discussion forums: Visual analysis with lda topic models. LOPAL '18 (2018)