グループディスカッション自動評価手法の実証研究

Empirical Study of an Automated Group Discussion Evaluation Method

増田　武史/Takeshi MASUDA・広瀬　啓雄/Hiroo　HIROSE

公立諏訪東京理科大学大学院工学・マネジメント研究科

[Abstract]
Low-cost individual evaluations of participants in group discussions are sought after in contexts such as human resource development, selection, and recruitment. This study aims to develop a method for automatically and multifacetedly evaluating discussions by proposing a method to structure discussions based on transcribed data and automatically annotate statements. In the experiment, a dataset of discussions by groups of business professional was used. The discussions were classified into a three-tiered structure, after which annotations representing the types of statements were added within each tier, and individual evaluations were conducted using evaluation formulas. As a result, it was found that participants in discussions could be evaluated multifacetedly, similar to traditional observation evaluations by assessors. Additionally, it was demonstrated that this method allows for continuous scoring, which cannot be expressed by the traditional five-point scale evaluations by assessors. By using this method, it becomes possible to conduct low-cost evaluations of discussions without relying on evaluators with advanced observation skills. In this paper, the annotations used in the individual evaluation of each speaker are manually assigned. In addition, the links between the referential relationships between the statements have also been manually assigned. Therefore, it is currently difficult to evaluate a large number of discussions. In the future, we would like to consider automating the process of annotating the categories of statements and assigning the reference relations to structure the statements into a graph.

[キーワード]
グループディスカッション、貢献度評価、自然言語、議論構造、アノテーション

1.　はじめに

1.1.　背景

企業における従業員の能力アセスメントや選抜、採用時の集団面接などでグループディスカッションが用いられている。ディスカッションにおける参加者個人の言動から、ペーパーテストでは評価が難しい現実世界でのパフォーマンスやリーダーシップを評価することを目的としている。このようなディスカッションの評価ではアセッサーが議論風景を観察して評価をすることが広く行われている。この観察評価を行う際には、公平性の観点から評価者のバイアスや評価のバラつきを極力排除しなければならない。そのためにはアセッサーに高度な熟練が要求される。また、一人のアセッサーが一度に観察できる人数が限られるため、大人数の評価を行うには多数のアセッサーを必要とする。このため、高度な技能を持つアセッサーの育成や確保と一度に処理できる人数の限界から、グループディスカッションの観察評価はコストのかかる評価方法となっている。そこで、グループディスカッションの評価をアセッサーの熟練技能に頼らずに経験の浅い評価者でもある程度客観的に行う方法が求められるようになった。

人が対面して議論や対話を行う状況の分析については、音声や身体的な表現をカメラやセンサーで取得するmultimodal learning analyticsが研究されてきた[1]-[3]。これによると参加者の感情も含めた高度な分析が可能になるが、カメラやセンサーの設置が必要なため会場設営が大掛かりになるため広く用いられている方法ではない。

同期的な口頭でのディスカッションではなく、オンラインフォーラムの書き込みを議論と見立てた研究も行われている。フォーラムのスレッドから重要発言を抽出する方法や、ディスカッション全体を要約する方法などが研究されてきた[4], [5]。しかし個人の能力を多面的に評価することはあまり行われてこなかった。

書き言葉である論文やエッセイの自動評価についての研究もなされている[6], [7]。エッセイは個人ごとに評価されるものの、多人数のインタラクションが発生している状況ではないため、対人能力の評価はできない。

従来は口頭でのディスカッションにおける参加者個人の評価をデータに基づいて行うためには参加者ごとにマイクを用意したり、録音した音声を文字起こししたりする必要があったが、それを簡易に行う方法がなかったため研究や実用化開発の対象になりにくかった。しかし近年急速にオンライン会議が普及し、容易にディスカッション参加者の音声を録音することができるようになってきている。また音声の自動文字起こしも実用的な精度になってきている。

1.2. 本研究の目的と意義

本研究では、口頭でのディスカッションを文字起こしデータに基づいて客観的かつ多面的に評価する方法を開発することを目的とする。

これにより、高度な評価技能を持つアセッサーの観察評価に頼ることなく、記録されたディスカッションのデータをもとにディスカッション参加者の評価を行うことができるようになる。また、従来は多人数のディスカッションを評価する際には複数のアセッサーが観察を行っており、評価の一貫性や公平性を保つためにアセッサー同士の調整が必要になる点でもコストがかかっていたが、自動評価システムを導入することにより評価の客観性や公平性が確保でき、かつ低コストで評価ができるようになると考える。多人数に対してディスカッション評価を導入することが従来よりも容易になるため、求める能力を持つ人材の発見にも寄与できると考える。

1.3. リサーチクエスチョンおよび本稿の構成

本稿では、以下のリサーチクエスチョンについて検討を行う。

リサーチクエスチョン

ディスカッション参加者の能力をディスカッションの音声記録の文字起こしに基づいて低コストかつ多面的に評価するためにはどうすればよいか?

これについて本論文では次の構成で検討を行う。まず2章で先行研究についてレビューする。次に３章でディスカッション評価のための構造化方法を定義し、ディスカッションの自動評価方法について述べる。４章で少数のディスカッションサンプルを用いた実験について述べ、5 章で実験から得られた結果を示す。最後に 6 章で本研究をまとめる。

2. 先行研究レビュー

議論を構造化する研究は古くから行われている。Kunz and Rittel[8]は 1970 年代に IBIS (Issue-Based Information Systems) を提案している。IBIS では議論を意思決定のために行うものとして、issue, position, argument などの論理を構成する要素に分解して構造化している。また、IBIS をベースにコンピュータ上で議論の構造をとらえた gIBIS[9], [10]をはじめとして議論を構造化して表すフレームワークには様々なものが提案されている[11], [12]。議論を構造化して表示するだけでなく、そこから議論の要約を作成する方法も研究されている[13]。

自然言語で表現された議論から構造を識別・抽出することは Argument mining とも呼ばれ様々な研究がなされている。Argument mining の対象となる議論は Lawrence and Reed[14]が指摘するように主張、根拠、反論、前提条件、結論などが含まれる。Alsinet et al.[15]はオンライン討論をコメント間の攻撃関係を意見の不一致度をモデル化した。

書き言葉であるエッセイを議論構造としてとらえモデル化する研究も行われている[16], [17]。Stab and Gurevvch[16]はエッセイ中の文章に対して主張と前提、主張に対する支持や攻撃を示すアノテーションを行い議論のモデル化を行った。

これらの先行研究はディベートのような明確な論点と対立構造がある議論や文書を研究対象としており、主張や反論などの発言の論理的な意味あいから議論を構造化しているものが多い。一方で本研究が対象とする人材育成や選抜、面接で用いられるグループディスカッションでは、必ずしも明確な論点や対立構造があるわけではない。これは Collaborative Discussion と呼ばれる[18]。協調的な対話のプロセスを通じて参加者個人が議論グループ全体に対してどのような貢献を行ったかについて本研究では評価を行う。それに対しては従来の議論の構造化に用いられる主張や反論などの論理を示すアノテーション以外の情報も必要となる。

発言のアノテーションについても様々な方法が試みられている[19]-[23]。また、一つ一つの発言ではなく議論中の参加者の役割を分類する方法も試みられている[24], [25]。Pianesi et al.[26]や藤本[27]は議論参加者をTask とSocio-Emotionalの2つのAreaそれぞれでfunctional roleを分類した。議論中の発言内容だけでなく、発言をあまりしないListenerや議論をまとめるSummarizerといった役割に着目した個人評価は現実社会での人物評価の場面においても重要である。

また従来は議論における発言を分析する際、意見の内容に注目し意思決定や結論、他者の発言への影響などが注目されてきた。Bartel[28]はグループのムードがパフォーマンスに影響することを指摘している。この点から、従来の議論分析はコンテンツとしての発言を重視している一方でディスカッショングループのムードについての検討が不十分であると言える。

非同期のオンラインフォーラムについての研究も様々に行われている

Klass[4]はオンラインフォーラムの発言と発言への返答の関係をノードとエッジからなるグラフ構造としてとらえ、重要発言を抽出する研究を行った。キーワード抽出にはグラフ構造内のノードの重要性を決定するためにTextRankモデルなどが用いられることがある[29]。村上[30]はオンラインフォーラムのスレッド構造から発言に有益発言指標値を与える方法を示している。

従来は同期的な口頭でのディスカッションと非同期の書き言葉からなるオンラインフォーラムは分析の手法も研究目的も異なるものであった。しかし現在では同期的な口頭でのディスカッションをリアルタイムでテキスト化することが技術的に可能になってきており、口頭でのディスカッションに対して非同期のオンラインフォーラムの分析手法を適用することも実用性があるようになってきている。

自然言語モデルを用いてテキストを分類することについても多くの研究がある。Mishev et al.[31]は機械学習による分類器やNLP トランスフォーマーなど複数のアルゴリズムを用いて経済ニュースの感情分析を行い、BERTをベースとしたトランスフォーマーモデルの有用性を示した。Devlin et al.[32]によって提案されたBERTはファインチューニングが容易であることから自然言語の分類によく用いられている。Prabhu et al.[33]は英語ではない言語で書かれたテキストの多クラス分類にBERT ベースのモデルを訓練して用い、良好なパフォーマンスが得られることを示した。

本研究では口頭でのディスカッションをテキスト化したデータをもとに議論参加者の個別評価を行う。現代では広く用いられているwebミーティングシステムの録音を文字起こししたものを用いるため、マイクやカメラ、センサーの設置が必要なく先行研究のディスカッション分析よりも実用性が高い。録音データのテキスト化や分類はNLPモデルを用い自動化した処理を行った後、手動によるチェック及び修正を加える半自動化処理とすることで、現在すべて人手で行っているディスカッション評価よりも低コストとなるようにする。また、論理性だけでない多面的な分析をディスカッションの発言データから行うことで、先行研究よりも広く総合的な観点での人物の能力評価ができ、選抜や育成に役立てることができる。

3. 方法

3.1. 議論の構造化

本研究では議論を発言のネットワークと見立て、発言をノード、発言と他の発言のつながりをエッジとしたグラフ構造で表す。発言グラフ構造を図1に示す。

他の発言を参照せず新たなクラスタを開始する発言を親発言，親発言から繋がる発言を子発言とする

図 1はAの発言をB・Cが参照し、Cの発言をE・F・Gが参照し、Gの発言をH・Iが参照する発言構造を示している。Iの発言はAの発言を直接参照した発言ではないが、間接的にAを参照した発言とみなすことができA発言の影響が及んでいると考える。

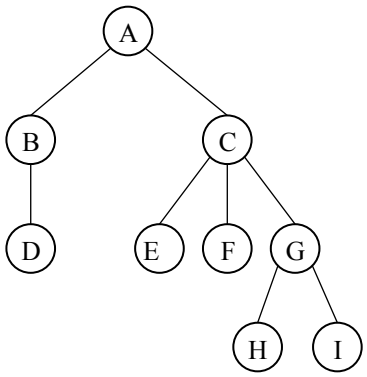

図 1　発言グラフ構造

ここで、本研究では高度な評価技能を持つアセッサーでなくともデータに基づいた評価が行えるようにするという目的から次の仮説を設定する。

仮説 1：発言内容の意味や重要性を解釈し吟味しなくとも、発言構造から発言の評価を行うことができる。

本研究では発言の評価について、村上[30]のオンラインフォーラムにおける有益発言を議論構造の返答関係から効率よく見つけることが可能になるという議論に基づき、口頭でのディスカッションにおいても発言を構造化することで同様に有益発言が発見できるという仮説 1 を設定した。
仮説 1 に基づいて本研究では多くの返答が得られた発言や、子発言が長く続く発言が高く評価される評価アルゴリズムとする。
図 1 の場合は B よりも C の発言の方が多くの返答が得られており、かつ子発言が長く続いているので高い評価を与える。ただし、すべての発言に対してこのアルゴリズムを適用すると、上流に位置する発言が必要以上に高く評価されることになるため、発言の連鎖に対して減衰率を適用することで距離が離れている発言の影響が大きくなり過ぎないようにする。

先行研究では議論の結論に影響を及ぼしている主張や意見、反論などにアノテーションを付与し構造化が行われている。そのため意見表明ではない相槌や冗談などの発言は評価の対象外であり、あらかじめ分析対象から除かれていた。一方で口頭でのディスカッションにおける発言は意見を出し合っているだけではなく、相槌をうったり、司会進行したり、冗談を言って場を和ませたり、意見以外の発言も多くみられる。
本研究では意思決定やディベートではない協調的議論の評価を行えるようにするという目的からディスカッション参加者の評価について、Bartel[28]によるグループのムードを決定する要因についての議論に基づき、感情的な相互作用を生み出すことに貢献した人物もディスカッションの価値を高める貢献があるという仮説を設定する。また、藤本[27]によるディスカッション中の役割分類の議論に基づき、議論を進行したり俯瞰して整理したりする人物もディスカッションへの貢献が評価できるという仮説とする。両者をまとめて次の仮説 2 とした。

仮説 2：発言の種類を構造上分けることによって、議論の本筋を構成する発言以外の多様な発言も評価することができる。

仮説 2 に基づいて、本研究では議論の発言構造を図 2 に示す通り 3 層で定義した。

図 2 に示す議論レイヤーは、評価者による議論評価の際にしばしば重視される項目から選定した。意見の表明や新しいアイデアの提示、それに対する補足や情報提供など議論の論理的な筋道を組み立てる発話が議論を構成する主な部分であるが、それ以外にも他者の発言を促したり、これまでの議論をまとめたりというファシリテーションが議論参加者の中で行われる。協調的議論ではあらかじめ司会者を決めるわけではないため、参加者の誰が自発的にファシリテーションを行うかはリーダーシップ能力を見極めたい評価者にとって関心が高い項目である。また、他者の発言に対して相槌を打ったり言葉で同意を示したりする行為は対人コミュニケーション能力を示すものとして評価される。

このことから議論レイヤーの３層を次のように定義する。

議論レイヤーの１つめはムードメイクレイヤー（Mood Creation Layer）である。協調的議論においてはよい雰囲気で議論が行えるように互いに協力してムードを作ることが重要である。他者が発言しやすいムードを作るような発言はこの層に位置する。例えば、他者の発言に対する同意や相槌、称賛、あいさつや冗談、笑い声などが該当する。

議論レイヤーの２つ目はディスカッションレイヤー（Discussion Layer）である。従来の議論分析で重視されてきた、議論の本筋となる論理を構築していく発言がこの層に位置する。例えば自分の主張の表明や新しいアイデアの提示、自分や他者の発言の補足や、他者に対する質問などが該当する。

議論レイヤーの３つ目はファシリテーションレイヤー（Facilitation Layer）である。議論を円滑に進めるために俯瞰して議論の進行を把握しファシリテーションを行う発言がこの層に位置する。例えば、他のアイデアはないかと議論を広げたり、議論を収束させたり、発言が少ない人に話を振ったりする発言が該当する。

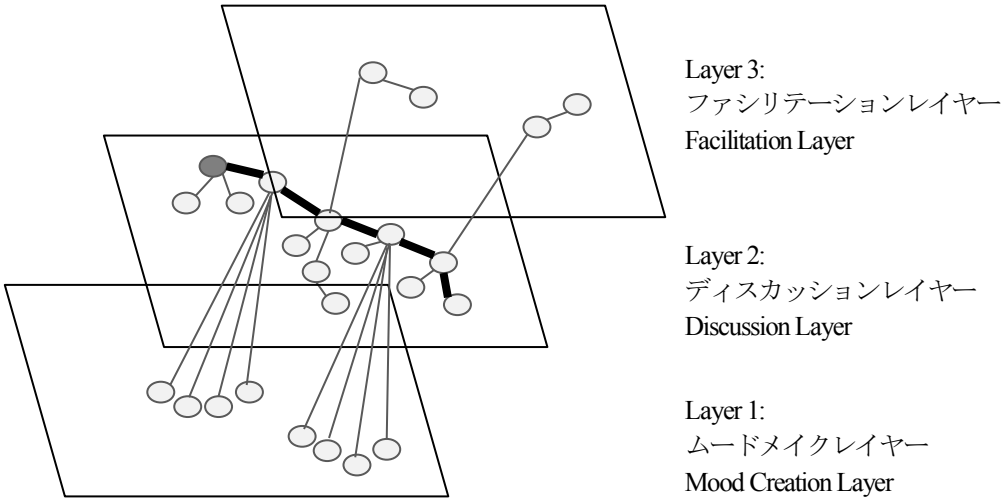

図 2　議論レイヤー

議論レイヤーに分けた発言に対して、さらに発言カテゴリのアノテーションラベルを付与する。本研究で用いる発言カテゴリのアノテーションラベルを表 1に示す

表 1　発言カテゴリ

議論レイヤー	発言カテゴリ	発言内容
ファシリテーション	Facilitation	議論の拡散、議論の収束、議論の要約、進行
ディスカッション	Opinion	主張、新しいアイデアの提示　など
	Argument Reinforcement	自分の発言に対する補足、他者の発言の補足
	Argument Development	既存の発言に関する確認、質問、質問への回答、関連する情報の提示
ムードメイク	Warm Up	あいさつ、相槌、冗談、笑い
	Agreement	他者の発言への同意
	Admiration	他社の発言への称賛、お礼

一つの発言に対して発言カテゴリは複数付与してもよいこととする。これは発言によっては口述前半でファシリテーション発言を述べた後に後半で自分の意見を述べるということもあるからである。

3.2. 発言の定量評価

ディスカッション発言をグラフ構造化し、3 層からなる議論構造に分解したうえで各発言にアノテーションを付与することで、発言が前の発言に対しての返答なのか、同意や相槌を示すもので新しい意見を加えているものではないのか、あるいは議論を俯瞰してまとめたり司会進行を行ったりするものなのか等の発言の特徴を分類することができる。

その上で、議論の本筋であるディスカッションレイヤーの発言に対して、前述の仮説 1 で検討した発言構造における位置づけによる定量的な発言評価を次のように行う。

・ディスカッションレイヤーの発言評価

ディスカッションレイヤーの発言は発言グラフ構造と発話カテゴリのアノテーションを用いて評価する。評価式は非同期ディスカッションにおける有益発言抽出を目的とした村上[30]の提示する評価式をもとに改良した式で、子発言が長く続く発言ほど評価点が高くなり、また多くの同意を集めたほど評価点が高くなるように設定する。

本研究で使用する評価数式を以下に示す。

$$cv_{(p)} = \left(sv_{(p)} + \sum_{i=0}^{N-1} cf_{(p_i)} \times cv_{(p_i)} \times af_{(p_i)} \right) \times cr_{(p)} \qquad (1)$$

ディスカッション中の発言 p に対する発言スコア$cv_{(p)}$は上記(1)式から算出される。

$sv_{(p)}$は自己発言スコアで、発言に与えられる初期数値である。

Nは自己発言を含む連鎖する発言の数である。

$cf_{(p)}$はカテゴリ係数で，子発言に付与された発話カテゴリにより数値が異なる。

$af_{(p)}$は減衰係数で，構造が深くなるにつれて影響を小さくするための係数である。

$cr_{(p)}$はカテゴリ確率を示し、発言が該当するカテゴリの重みを示す。一つの発言にはアノテーションによって複数の発話カテゴリが付与される可能性がある。例えば，ファシリテーションとディスカッションレイヤーの意見の 2 つの発話カテゴリが付与された発言であれば、各カテゴリの重み$cr_{(p)}$を掛けることでファシリテーションスコアとディスカッションレイヤーの意見スコアを算出する。これにより一つの発言が複数のレイヤーの側面を持っている場合もそれぞれのカテゴリごとのスコアを算出することが可能になる。

本稿では議論の本筋ではないファシリテーションレイヤーとムードメイクレイヤーの発言に対しては以下の方法で評価を行う。

・ファシリテーションレイヤーの発言評価

このレイヤーに該当する発言の回数を評価スコアとする。

・ムードメイクレイヤーの発言評価

このレイヤーに該当する発言の回数を評価スコアとする。

これにより、各発言は発言者、発話カテゴリ、評価スコア（発話カテゴリが複数付与されている場合は、複数の発話カテゴリとそれぞれの発話カテゴリに対する評価スコア）の属性を持つ。

3.3. 議論参加者別の評価の作成

算出した評価スコアを参加者別、発話カテゴリ別に積算する。さらにそれを議論レイヤーごとに標準化（偏差値化）する。これによって議論参加者の各レイヤーにおける発言を評価することができる。

最終的に表 2に示す評価ディメンジョンで個人評価結果を示す。この評価ディメンジョンはアノテーションを行った発話カテゴリを一部組み合わせたものである。

表 2 評価ディメンジョン

評価項目	概要
Opinion	主張や新しいアイデアを提示した発言がディスカッションに与えた影響の強さを示す
Argument Reinforcement	他者や自己の発言を補足・強化する力を表す
Argument Development	質問や発言内容の確認などを通じて議論を発展させる力を示す
Mood Creation	他者に共感したり称賛したり冗談を言ったりすることで発言しやすい雰囲気を作る力を示す
Facilitation	議論を俯瞰し発言を促したりまとめたりする力を示す

以上の方法により、本研究では発言の返答関係の構造と発言のレイヤー別アノテーションを用いて議論参加者の個別評価を行うことができる。

本研究の方法によると、高度なアセスメント技能を用いて議論における発言内容を吟味する必要がなく、簡易なアノテーションで評価を定量的に表すことができる。

また本研究のアノテーション方法によると、結論や主張が明確ではない協調的議論においても参加者を多指標に評価することができる。

3.4. ディスカッション評価の自動化

本研究の目的である、ディスカッション評価を低コストで行うための自動化方法について述べる。

従来の評価者による観察評価ではディスカッション参加者 4-6 名につき 1 名の評価者がディスカッションに同席し参加者の観察記録を取っていた。そののち記録を元に参加者個別の評価を行い、複数のグループがある場合はグループ間での評価の調整を行う。また評価者が複数人いる場合は評価者同士で評価基準のすり合わせも必要であった。このためディスカッション評価は高コストになっている。

このディスカッション評価を低コスト化するために、図 3に示すディスカッション自動評価システムを提案する。システムの構成は次の通りである。

1： ディスカッションシステム（Web ミーティングシステム）を用いてグループディスカッションを行う。ここで参加者の発言が録音される。
2: 文字起こしシステムによって録音データをテキストデータ化する。
3: ディスカッション構造化システムによりテキスト化されたディスカッションに議論レイヤー構造や発言カテゴリ等のアノテーションを付与する。
4: 個別評価システムにより構造化したディスカッションデータをもとに各発言に対して評価点を付与する。発言者別、評価観点別に評価点を集計し、ディスカッション参加者個別に多面的な評価を算出する。
5: ディスカッション自動評価システムの出力として、参加者個別の評価結果を返す。

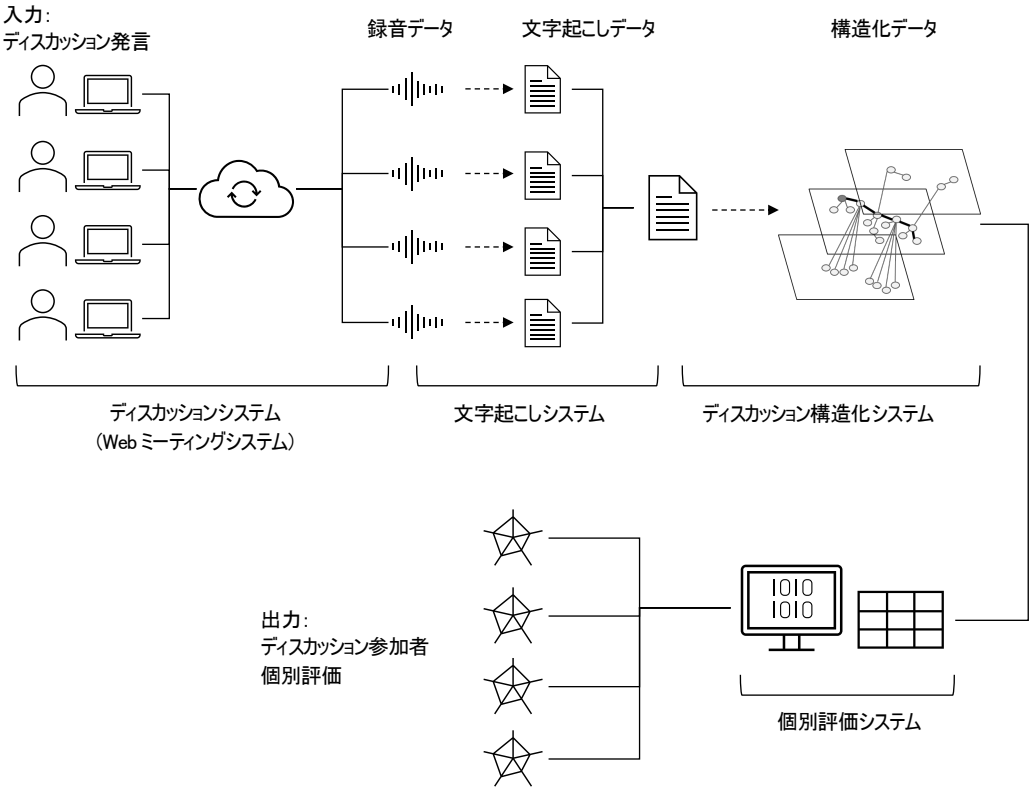

図 3　ディスカッション自動評価システム

4.　実験

4.1.　データセット

本実験で使用する議論データについて述べる。

ビジネスパーソンのべ16組によるディスカッションの録音記録を用いた。議論は例えば“テレワークで仕事をすることのポジティブな側面について考える””架空の企業を題材に出店戦略を検討する”のようなテーマで、5-6 人が 1 組となって行われた。ディベートのように賛成反対に分かれて議論を戦わせるのではなく、ディスカッションの中で参加者に立場の違いがなく、協力しながら検討を進めグループとしての結論を導く協調的議論のテーマ設定となっている。このような協調的議論は企業内の人材育成や採用時のグループディスカッションで多く用いられる形式であるため、本実験でも協調的議論のテーマを採用することとした。

議論は web 会議システムの zoom を用いて行い、発言を録音した。録音した議論データは文字起こし機能を用いてテキストデータに変換し、その後録音データと聞き比べて手動で修正を行った。テキストデータへの変換に際しては音声として聞き取れる発言はすべてテキストデータ化した。

得られたデータは、発言番号、タイムコード、発言者、発言の合計 4 種類の要素で、総発言数は5016回となった。データ化したディスカッション音声記録の一部を表 3 に示す。

表 3　データ化したディスカッション音声記録の抜粋

番号	タイムコード	発言者	発言
100	06:03.25	B	個人と、経済と、となんかそういう整理にするとか。ま、例えばですけど。
103	06:15.19	A	それいいですね。あの、個人の観点と、ま、社会全体なのか、っていう観点と。
104	06:19.03	F	うん
105	06:20.01	D	うんうん
112	06:30.12	D	そうすると個人は、そうすると
113	06:35.21	E	ワークライフバランスと

4.2. ディスカッション発言のレイヤー自動分類

議論データの一部を用いて発言の自動レイヤー分類の訓練データを作成した。ディスカッション発言に手動で3 種類のレイヤーの正解ラベルを付与したものを訓練用データと検証用データに分割した。作成したデータセットを表 4に示す。

表 4 レイヤー自動分類訓練用データセット

	データ数		
議論レイヤー	**訓練用データ**	**検証用データ**	**合計**
ファシリテーション	171	43	214
ディスカッション	848	212	1060
ムードメイク	992	248	1240

自動分類は日本語でプリトレーニングされた大規模自然言語モデルの BERT[34]と simpletransformers package[35]を用いプログラム化して行った。訓練データを用いてファシリテーション、ディスカッション、ムードメイクの 3 クラス分類の学習を行い、検証用データでモデルの妥当性を確認した。モデルのパラメータは simpletransformers のデフォルトを用いエポック数のみを設定した。

自動分類モデルを作成したのち、モデルの学習に用いていない新たなデータセットを用いてレイヤー自動分類の性能テストを行った。新たなデータセットには表 5に示す3回のディスカッションデータを用いた。

表 5 性能検証用テストデータ

	データ数		
議論レイヤー	**議論 1**	**議論 2**	**議論 3**
ファシリテーション	39	38	43
ディスカッション	547	484	556
ムードメイク	213	99	305

4.3. ディスカッション参加者の個人評価

テキスト化した議論データのうち2回分の議論データ（議論a、議論b）を用いて議論参加者の個人評価用実験データを作成した。発言をグラフ構造化するために、1 回の発言を一つのノードとし、発言間の参照関係を示すエッジの付与を行った。エッジの付与はある発言がどの発言を参照しているかを発言番号で指定することによって行った。今回の実験では作業はディスカッションのアセッサー経験を持つ有識者が手作業で行った。

次に発話カテゴリのアノテーション作業を行った。テキストデータ化したディスカッション中の発言に対して、表1に示した発話カテゴリを手動で付与した。まずディスカッションのアセッサー経験を持つ有識者1名が研究補助者2名に対してアノテーション作業指導を行い、その後合計3名が別個に作業を行った。その結果、2回のディスカッションの合計423発言にアノテーションが行われた。

3 名中 3 名が同じアノテーションラベルを付与した発言は全体の 41%で、3 名中 2 名が一致した発言は全体の48%であった。合わせて全発言数の 89%は3名中2名以上のアノテーション付与が一致している。

アノテーションの一致率はそのままカテゴリ確率の変数として用いた。例えばある発言に対して、3 名中 3 名が「ファシリテーション」のアノテーションを付与し、さらに3名中2名が「意見」のアノテーションを付与した発言は、ファシリテーションのカテゴリ確率が 1、意見のカテゴリ確率が 0.67 となる。

ファシリテーション、ディスカッション、ムードメイクのどの層にも属さない発言は評価対象外とした。

アノテーションの後、発言の定量評価の項で示した計算式に従って各発言の評価スコアを算出した。

5. 結果

5.1. ディスカッション発言のレイヤー自動分類

表 6にレイヤー分類結果の正解率を、**エラー！参照元が見つかりません。-エラー！参照元が見つかりません。**に再現率、適合率、F1 スコアを示す。

表 6　レイヤー分類モデルの正解率

性能検証用議論データ	正解率
議論 1	0.87
議論 2	0.84
議論 3	0.82

表 7　性能検証用議論のレイヤー自動分類結果

議論レイヤー	再現率			適合率			F1 スコア		
	議論 1	議論 2	議論 3	議論 1	議論 2	議論 3	議論 1	議論 2	議論 3
ファシリテーション	0.31	0.29	0.23	0.35	0.24	0.33	0.33	0.27	0.27
ディスカッション	0.92	0.87	0.93	0.92	0.95	0.83	0.92	0.90	0.88
ムードメイク	0.84	0.94	0.72	0.82	0.70	0.85	0.83	0.80	0.78

表 6に示したよう大規模自然言語モデルのBERTと訓練データを用いて学習を行ったレイヤー自動分類モデルを用いて80%以上の正解率で発言を3レイヤーに分類することができた。ただし、**エラー！ 参照元が見つかりません。**に示すようにレイヤーごとの性能を見るとディスカッション・ムードメイクレイヤーはF1スコアが0.78–0.92と高精度で分類できているのに対して、ファシリテーションレイヤーのF1スコアは0.27–0.33と低い値にとどまった。

5.2.　ディスカッション参加者の個人評価

評価スコアの算出に用いたパラメータを表 8に示す。

パラメータの値は村上[30]の先行研究を参考に設定した。議論のメインストーリーを構成する発言である意見（発話カテゴリ：Opinion）を基準の「1」とし、その意見に他者からの賛同や意見補強が多く集まるほどスコアが高くなるように設定した。短い発話となることが多いムードメイクの係数は相対的に小さくしている。

表 8　パラメータ

パラメータ	議論レイヤー	発話カテゴリ	設定値
$sv_{(p)}$			0.5
$cf_{(p)}$	Facilitation	Facilitation	0.75
	Discussion	Opinion	1.0
		Argument Reinforcement	1.25
		Argument Development	0.75
	Mood Creation	Warm Up	0.5
		Agreement	1.0
		Admiration	1.25
$af_{(p)}$			0.3

実験の結果得られた評価スコアを表 9、表 10に示す。表中のA–Fは議論参加者を示す。

表 9　議論a　本研究手法による評価スコア

評価ディメンジョン	A	B	C	D	E	F
Opinion	45.19	69.12	37.22	48.26	64.83	50.10
Argument Reinforcement	40.90	62.99	37.22	56.24	62.99	46.42
Argument Development	38.44	57.47	37.83	52.56	48.88	43.35
Mood Creation	38.15	48.61	38.15	50.17	60.63	64.29
Facilitation	39.27	58.21	37.38	56.31	63.89	44.95

表 10　議論b　本研究手法による評価スコア

評価ディメンジョン	A	B	C	D	E	F
Opinion	54.51	74.10	37.72	50.31	54.51	48.91
Argument Reinforcement	46.11	65.70	37.72	44.71	58.71	44.71
Argument Development	43.31	57.31	37.72	41.91	41.91	60.11
Mood Creation	43.30	49.18	35.95	51.14	51.63	68.79
Facilitation	52.93	58.78	29.51	52.93	58.78	47.07

本研究で使用したディスカッションデータを構造化した発言グラフ構造の一部を図 4 に示す。ノードの大きさが各発言の評価スコアを表す。作図にはPython ライブラリの NetworkX を使用した。

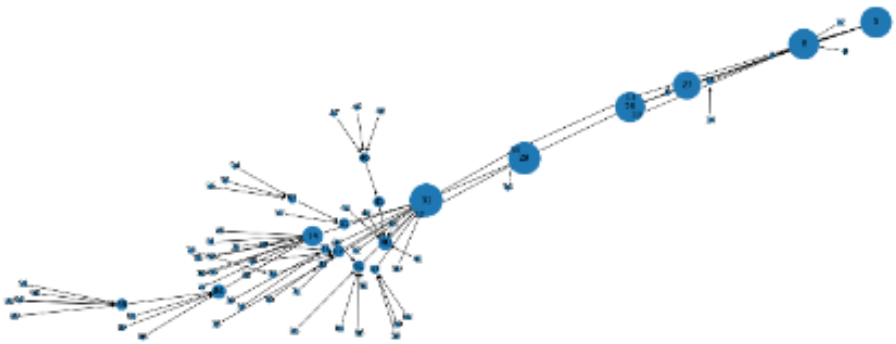

図 4　可視化した発言グラフ構造

図 4 に示されているように子発言が多い、あるいは子発言が長く続く発言の評価スコアが高くなっている。

比較のために同じディスカッションデータに対して、従来手法であるアセスメント技能を持つ有識者による評価を実施した。アセッサーによる評価は5段階で行った。評価基準を表 11 に、評価結果を表 12、表 13 に示す。

表 11　アセッサーによる5段階評価

評価スコア	基準
5	議論に大きく貢献している
4	議論に貢献している
3	どちらでもない
2	議論への貢献が少ない
1	議論への貢献が見られない、もしくは発言がない

表 12　議論 a アセッサーによる評価スコア

評価ディメンジョン	A	B	C	D	E	F
Opinion	3	5	1	3	4	4
Argument Reinforcement	4	4	1	3	5	2
Argument Development	1	3	1	3	3	2
Mood Creation	4	4	2	4	4	5

Facilitation	2	3	1	3	4	2

表 13 議論b アセッサーによる評価スコア

評価ディメンジョン	A	B	C	D	E	F
Opinion	3	5	1	3	4	3
Argument Reinforcement	3	4	1	3	4	3
Argument Development	2	3	1	2	2	2
Mood Creation	4	4	1	4	4	5
Facilitation	3	4	1	2	4	2

本研究手法と従来手法の評価スコアを評価ディメンジョンごとに相関係数を求めた結果を表 14 に示す。

本研究手法による評価と従来の評価者による観察評価に高い相関が認められる。今回の実験において採用した評価ディメンジョンにおいては、本研究手法による評価はアセッサーによる観察評価と同様にディスカッション参加者を多面的に評価できている。

表 14 本研究手法と従来手法の相関

評価ディメンジョン	相関係数
Opinion	0.91
Argument Reinforcement	0.76
Argument Development	0.77
Mood Creation	0.76
Facilitation	0.91

6. 考察と含意

6.1. 考察

本研究手法により高度なアセスメント技能を持つ評価者でなくとも、議論の発言構造と議論レイヤーごとに付与したアノテーションを用いて議論参加者を多面的に評価できることが事例により示された。

従来の議論分析では主張や結論が明確なディベートや意思決定のための議論を対象とした構造化手法が提案されていた。それに対して本研究では議論の論理構造を明らかにする目的以外に、意見が出しやすいムード作りや議論を円滑に進めるファシリテーションへの貢献度合いも評価できる議論レイヤー構造を発見した。今回の実験ではディスカッションレイヤーのみ各発言の評価点を計算式によって求め、ファシリテーションレイヤーとムードメイクレイヤーは発言回数によって評価を行った。ディスカッションレイヤー以外は簡易な評価方法であるが、それでも従来手法である評価者による観察評価とおおむね同様の結果が得られている。本研究手法で提案するように議論を複数の層に分けることで各層ごとに異なった評価アルゴリズムを適用することも可能になる。これにより、評価目的に応じた柔軟な評価が行えるようになるため、実際の運用場面において実用的な評価方法となっている。

本研究手法では発言の相互参照関係に着目して評価点を与えており、発言の言葉の意味は考慮していない。一方で観察評価を行う評価者は発言者の言葉から意図をくみ取り評価を行うことが通常である。この点で本研究手法と従来手法である観察評価は評価対象が大きく異なるとも考えられる。にもかかわらず今回の評価ディメンジョンにおいては両者に高い相関が認められたことは注目に値する。発言内容の分析にまで踏み込まずとも、発言構造から連鎖関係を分析することで、ディスカッション参加者同士で発言意図が共感され議論の盛り上がりに貢献したものが自然と高く評価されたためと考えられる。しかし、発言の内容そのもの（例えば知識の豊富さや語彙の高度さ、回答の妥当性など）を評価する場合は本研究の特徴である発言の連鎖関係で評価を行うことは困難であることが予想される。この点から本研究手法は相互にやり取りが行われるディスカッションにおける貢献度の評価に特化した手法であると考えられる。

また、従来の議論分析では論理的な議論に適したアノテーションが付与されていたが、本研究では意見の対立構造がない議論や明確な結論に導かれない協調的議論にも適用できるアノテーションを提案している。このアノテーションを用いて算出した評価スコアを偏差値化することで、従来よりも分解能の高い評価が可能になってい

る。従来は人間が観察して評価を行うため5段階や7段階評価のような段階評価が行われていた。評価点が階段状となるため、例えば3点と4点の間の点が付けられないとか、5段階の5以上の点が付けられないなどの制約が生じていた。しかし本手法では計算式によって求めた評価スコアを偏差値化して表示するため連続値を取ることができる。これによって人間による評価ではできなかった細かい差をつけることが可能になっている。

今回の実験では発言のレイヤー分類を大規模自然言語モデルの BERT を用いて行ったが、特にファシリテーションレイヤーの分類精度が低い結果となった。その原因としてはディスカッション中のファシリテーション発言がそもそも少ないことと、ディスカッション発言とファシリテーション発言の内容が似ていることがあるためと考える。ファシリテーションレイヤーとディスカッションレイヤーの分類精度を高めるために、今後さらにデータ数を増したうえで分類モデルの訓練を行うことと、異なる自然言語モデルを用いた自動分類モデルを作成することで精度の向上を狙いたい。

本稿では発言者の個別評価で用いたアノテーションは手動で付与したものを用いている。また、発言同士の参照関係のリンク付与も手作業で行っている。本研究の狙いが評価作業のコスト低減と処理能力の向上であるため、この点ではまだ改善の余地がある。従来手法のように発言内容を吟味して評価を行うアセスメント技能ほどのスキルは必要とせず、簡易なアノテーション作業で評価が行えるため従来手法ほどコストがかかるわけではないが、手動作業が残っている部分がボトルネックとなり短時間で大量のディスカッションを評価することは現時点では困難である。そのため今後はさらに自動化できる範囲を広げることを検討する。自動化の対象としては発言カテゴリのアノテーション付与、発言をグラフ構造化するための参照関係の付与を想定しており、自然言語モデルと深層学習を用いて行えるか可能性を検討したい。

6.2. まとめと今後の展望

本研究では、口頭でのディスカッションを文字起こしデータに基づいて低コストかつ多面的に評価する方法を開発することを目的として、協調的議論への貢献度を定量的に評価する議論の構造化及びアノテーション方法を提案した。ディスカッション中の発言を3層の議論構造に分類しさらに各レイヤー内で協調的議論に適したアノテーションを付与したうえで評価式を用いて評価を行うことで、議論参加者を多面的に評価することが可能であることが分かった. また、本研究手法を用いることで従来の評価者による段階評価では表現できない連続的なスコアリングも可能になることが示された。高度な評価技能を持つアセッサーの観察評価に頼ることなく、記録されたディスカッションのデータをもとにディスカッション参加者の評価を行うことで客観性や公平性が確保でき、かつ低コストで評価ができるようになると考える。

今後は現在手動で行っている参照関係の付与や発言カテゴリのアノテーション付与の自動化を検討する必要がある。自動化により手動で行っている部分が少なくなれば、多人数に対してディスカッション評価を導入することが従来よりも容易になるため、社会での運用場面では人材育成や選抜、評価、採用などで求める能力を持つ人材の発見にも寄与できると考える。

[参考文献]

[1] S. Okada *et al.*, “Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets,” in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, Tokyo, Japan, 2016, pp. 169-176.

[2] U. Avci and O. Aran, “Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction,” *IEEE Trans. Multimedia*, Vol. 18, No. 4, pp. 643-658, Apr. 2016.

[3] G. Murray and C. Oertel, “Predicting Group Performance in Task-Based Interaction,” in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, Boulder, CO, USA, 2018, pp. 14-20.

[4] M. Klaas, “Toward indicative discussion fora summarization,” *UBC CS TR-2005*, Vol. 4, 2005.

[5] B. Sharifi, M.-A. Hutton, and J. K. Kalita, “Experiments in Microblog Summarization,” in *2010 IEEE Second International Conference on Social Computing*, 2010, pp. 49-56.

[6] E. L. Snow, L. K. Allen, M. E. Jacovina, S. A. Crossley, C. A. Perret, and D. S. McNamara, “Keys to Detecting Writing Flexibility Over Time: Entropy and Natural Language Processing,” *Learning Analytics*, Vol. 2, No. 3, pp. 40-54, 2015.

[7] Z. Ke and V. Ng, “Automated essay scoring: A survey of the state of the art,” in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China, 2019.
[8] W. Kunz and H. W. J. Rittel, “Issues as elements of information systems,” *Institute of Urban and Reigional Development, University of California at Berkeley Working Paper*, Vol. 131, p. 14, 1970.
[9] J. Conklin and M. L. Begeman, “gIBIS: a hypertext tool for exploratory policy discussion,” *ACM Trans. Inf. Syst. Secur.*, Vol. 6, No. 4, pp. 303-331, Oct. 1988.
[10] K. C. Burgess Yakemovic and E. J. Conklin, “Report on a development project use of an issue-based information system,” in *Proceedings of the 1990 ACM conference on Computer-supported cooperative work*, Los Angeles, California, USA, 1990, pp. 105-118.
[11] T. van Gelder, “Enhancing Deliberation Through Computer Supported Argument Visualization,” in *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, P. A. Kirschner, S. J. Buckingham Shum, and C. S. Carr, Eds. London: Springer London, 2003, pp. 97-115.
[12] A. Selvin *et al.*, “Compendium: Making meetings into knowledge events,” 2001.
[13] E. Barker and R. Gaizauskas, “Summarizing Multi-Party Argumentative Conversations in Reader Comment on News,” in *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, 2016, pp. 12-20.
[14] J. Lawrence and C. Reed, “Argument mining: A survey,” *Comput. Linguist. Assoc. Comput. Linguist.*, Vol. 45, No. 4, pp. 765-818, Jan. 2020.
[15] T. Alsinet, J. Argelich, R. Béjar, and S. Martínez, “Measuring user relevance in online debates through an argumentative model,” *Pattern Recognit. Lett.*, Vol. 133, pp. 41-47, May 2020.
[16] C. Stab and I. Gurevych, “Annotating Argument Components and Relations in Persuasive Essays,” in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 1501-1510.
[17] C. Stab and I. Gurevych, “Parsing argumentation structures in persuasive essays,” *Comput. Linguist. Assoc. Comput. Linguist.*, Vol. 43, No. 3, pp. 619-659, Sep. 2017.
[18] R. K. Sawyer, “Creative Teaching: Collaborative Discussion as Disciplined Improvisation,” *Educ. Res.*, Vol. 33, No. 2, pp. 12-20, Mar. 2004.
[19] H. Bunt, V. Petukhova, D. Traum, and J. Alexandersson, “Dialogue act annotation with the ISO 24617-2 standard,” in *Multimodal Interaction with W3C Standards*, Cham: Springer International Publishing, 2017, pp. 109-135.
[20] S. Renals, T. Hain, and H. Bourlard, “Recognition and understanding of meetings the AMI and AMIDA projects,” in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2007, pp. 238-247.
[21] A. Fang, J. Cao, H. Bunt, and X. Liu, “The annotation of the Switchboard corpus with the new ISO standard for dialogue act analysis,” in *Workshop on Interoperable Semantic Annotation*, 2012, p. 13.
[22] J. Sinclair and M. Coulthard, *Towards an analysis of discourse*. taylorfrancis.com, 2013.
[23] “ISO 24617-2:2020,” *ISO*, 2020. [Online]. Available: https://www.iso.org/standard/76443.html. [Accessed: 21-Jun-2024].
[24] K. D. Benne and P. Sheats, “Functional roles of group members,” *J. Soc. Issues*, Vol. 4, No. 2, pp. 41-49, 1948.
[25] A. J. Salazar, “An Analysis of the Development and Evolution of Roles in the Small Group,” *Small Group Research*, Vol. 27, No. 4, pp. 475-503, Nov. 1996.
[26] F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti, “A multimodal annotated corpus of consensus decision making meetings,” *Language Resources and Evaluation*, Vol. 41, No. 3, pp. 409-429, Dec. 2007.

[27] M. Fujimoto, “Team Roles and Hierarchic System in Group Discussion,” *Group Decision and Negotiation*, Vol. 25, No. 3, pp. 585-608, May 2016.
[28] C. A. Bartel and R. Saavedra, “The Collective Construction of Work Group Moods,” *Adm. Sci. Q.*, Vol. 45, No. 2, pp. 197-231, Jun. 2000.
[29] R. Mihalcea and P. Tarau, “TextRank: Bringing Order into Text,” in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004, pp. 404-411.
[30] 村上明子, “オンラインディスカッションにおける有益発言の抽出,” *言語処理学会第 14 回年次大会, 2008*, pp. 352-355, 2008.
[31] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, “Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers,” *IEEE Access*, Vol. 8, pp. 131662-131682, 2020.
[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171-4186.
[33] S. Prabhu, M. Mohamed, and H. Misra, “Multi-class Text Classification using BERT-based Active Learning,” *arXiv [cs.IR]*, 27-Apr-2021. [Online]. Available: https://arxiv.org/pdf/2104.14289 [Accessed: 18-Jun-2024].
[34] Tohoku University, *bert-japanese: BERT models for Japanese text*. Github. [Online]. Available: https://github.com/cl-tohoku/bert-japanese [Accessed: 21-Jun-2024].
[35] T. Rajapakse, *simpletransformers*. Github. [Online]. Available https://github.com/ThilinaRajapakse/simpletransformers. [Accessed: 21-Jun-2024].

(2024 年 8 月 25 日受理)