修 士 論 文

複数人議論の品質評価に向けた
コーパスおよび評価モデルの構築

指導教員: 嶋田 和孝　教授

九州工業大学大学院情報工学府
先端情報工学専攻

2020 年度

塩田 宰

論　文　概　要

九州工業大学大学院情報工学府 先端情報工学専攻 知能情報工学専門分野

学生番号	18676011	氏　名	塩田 宰
論文題目	複数人議論の品質評価に向けた コーパスおよび評価モデルの構築		

1　はじめに

近年，大学入試や就職試験において受験者の知識以外の能力（非定形知）を測ることが一般的となってきている．よく用いられる手法としてグループディスカッションがあり，試験官はグループの活動を観察しながら各受験者を評定する．しかし，正解/不正解の存在しない課題に取り組む様子を評価することは多大な労力を要する．加えて，評価者が常に定量的で客観性の高い評価を行っているとは限らない．そのため，議論参加者の発言内容や表情などを基に，議論の品質を採点・フィードバックするシステムにより評価者の活動を支援することは重要な課題の 1 つである．そこで，本研究は複数人議論と品質ラベルがペアとなった議論コーパスの作成，および複数人議論自動評価モデルの構築を目指す．

2　コーパスの構築

本研究では「討論」と「合意形成」という性質の異なる 4 人一組のグループによる議論を計 10 対話（5 グループ分）収集した．その後，収集した映像・音声データを基に，人手による発話の書き起こし，OpenPose[1] および OpenFace[2]による各議論参加者の骨格・顔特徴点，Surfboard[3]による音声特徴量を獲得し，マルチモーダル複数人議論コーパスを構築した．また，各対話に対してトピックに基づくセグメンテーションを実施し，10 対話のデータを合計 178 個の議論セグメントに分割した．その後，議論の品質評価の理論 [1] を基に，議論の合理性（Re）や有効性（Ef）に関する計 10 種類の軸で各議論セグメントを評定した．議論の合理性とは，議論が議題と関連しているかなど，議論の内容に関連した評価軸である．議論の有効性とは，聴衆がその議論を信用できるかや，議論のわかりやすさなど，議論の受け取り手の感情に関連した評価軸である．

3　提案手法

本研究における議論の品質推定は，ある議論セグメントの発話系列の情報を入力として，その議論セグメントに付与されたスコアを推定することである．本研究では発話の言語，表情，顔，音声の情報を BERT[4]，OpenPose，OpenFace，Surfboard によってベクトル化し，それぞれを SVM，LSTM，注意機構付き LSTM，階層型 LSTM へ入力し，議論セグメントのスコアを推定する．

表 1: 各設定で最高性能となったモデルの評価値

合理性（**Re**）				
モダリティ	L	M	H	Ave.
unimodal	0.000	0.611	0.340	0.451
bimodal	0.000	0.535	0.446	**0.459**
multimodal	0.000	0.569	0.306	0.415
有効性（Ef）				
モダリティ	L	M	H	Ave.
unimodal	0.000	0.580	0.352	0.459
bimodal	0.000	0.598	0.376	0.478
multimodal	0.000	0.627	0.365	**0.490**

4　実験

表 1 に各モダリティの組み合わせにおいて最大値となったスコアについてまとめる．合理性の評価においては議論の内容を評価するため，入力情報のモダリティを拡張しても評価性能の向上は見込めないことを確認した．一方，議論の信用性や明瞭性を評価する有効性においては発言者および聴衆の言語外の動作を考慮すると評価性能が向上し，特に，全てのモダリティを入力としたときに推定性能が最大となることを確認した．

5　おわりに

本研究は議論の自動品質評価を目的として，マルチモーダル複数人議論コーパスと自動評価モデルを構築した．今後は対話数の拡充を行うと共に，各評価軸の特性を考慮したモデルの構築による評価性能の向上が課題としてあげられる．

参考文献

[1] H. Wachsmuth et al. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of EACL*, pp. 176–187, 2017.

[1]https://github.com/CMU-Perceptual-Computing-Lab/openpose

[2]https://github.com/TadasBaltrusaitis/OpenFace

[3]https://github.com/novoic/surfboard

[4]https://github.com/cl-tohoku/bert-japanese

目 次

第1章　はじめに

近年，明確な答えのない課題に対して学習者同士で議論を進め，協力しながら問題解決に取り組む問題基盤型学習（Problem-Based Learning）や協調学習（Collaborative Learning）が教育において注目を集めている．また，大学入試や就職試験においてもグループディスカッションを用い，座学を通じて獲得可能な形式知（hard skills）だけでなく，受験者のコミュニケーション能力や性格特性など，他者と交流する際に影響する能力である非定形知（soft skills）を評価の指標に導入することが一般的になってきている．これらの背景には，経済・社会活動においてテストの成績やIQで測られるような形式知だけでなく，非定形知に関する能力も同様に重要視する必要があるとされてきたからである [1]．非定形知は各人の個性などに依存するものの，気づきや内省を繰返すことで向上，改善することが知られており，形式知と同様に非定形知についても教育を行っていくことは重要である．

教育において非定形知を訓練する問題基盤型学習の実施方法の1つに討論や合意形成を行うグループディスカッションがある．例えば,「小学校におけるプログラミング教育の必要性の是非と具体案・改善案についてまとめてください」という題材をグループに与え，その活動の様子を評価者が観察・評価を行い，グループ・個人それぞれにフィードバックを行う学習方法である．学校現場の教師がこの学習方法を導入する場合，1つのクラスにディスカッションを実施するグループが同時に複数存在することになり，全てのグループ・個人の能力や成果を様々な角度で評価・フィードバックを行うことは多大な労力を要する．更に，ディスカッションでは正解のない課題を取り扱うため，グループや個人に対して導き出される評価が常に定量的で客観性が保証されているとは限らない．そこで，グループの様子を様々なモダリティの情報（e.g. 発言内容，動作，音声）から獲得して議論内容の要約や評価を可視化するシステムを構築することができれば，評価者の評価活動の負担を軽減することができると考えられる（図1.1）．また，議論中や議論後に学習者自身が議論について細かく振り返ることは困難であるが，このようなシステムによって議論内容が記録されていれば，活動に関する振り返りを容易に行えるようになる．

非定形知の評価は先ほども述べたように，正解/不正解を採点する評価活動より

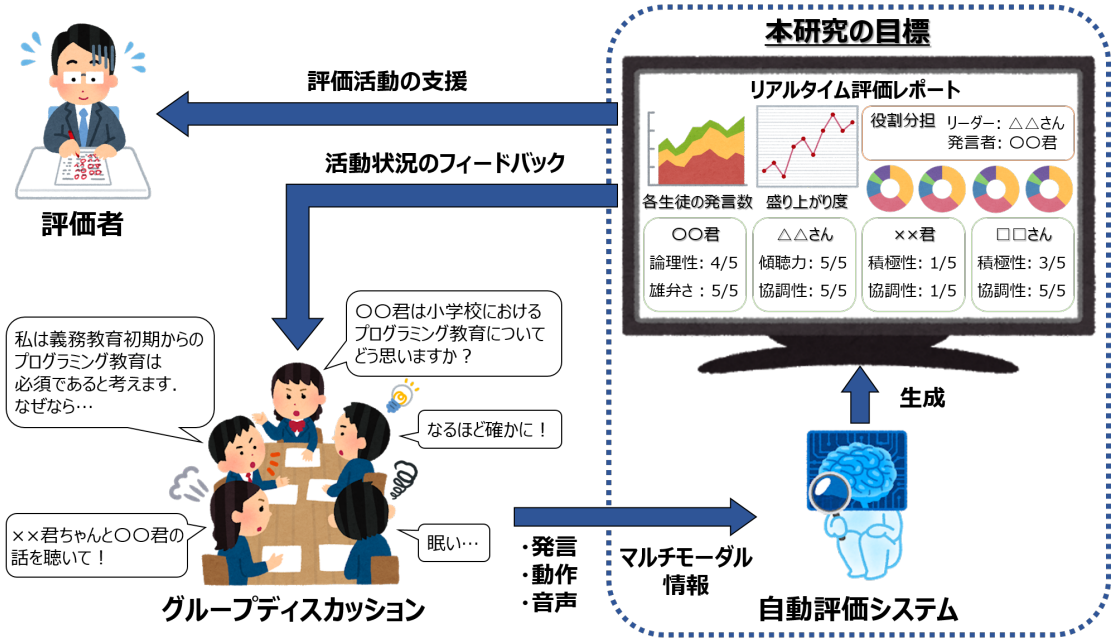


図 1.1 研究概要

も高難度な作業であり評価者の負担が大きい．そのため，非定形知の評価を自動で行うシステムの実現は古くから関心が持たれている．自然言語処理の分野において教育応用を目的とした議論の品質評価の研究の例としては Automated Essay Scoring（AES）[2] などがある．しかしながら，話し言葉による陳述の評価は書き言葉の論述の評価よりも困難な点が複数ある．例えば，エッセイは論述の内容を熟考できるため，文章がトゥルーミンのモデル [3] に代表されるような議論構造で整っており，かつ他者からのリアルタイムな意見や質問を考慮する必要がない．一方で，対面形式の議論の場合，陳述時間は短時間であり，話し言葉や言い淀みを含む場合や論理構成が崩れることもしばしばある．また，議論に参加している他者からのリアルタイムな質疑や反駁といった外的要因も議論の質に大きな影響を及ぼすと考えられる．さらに，書き言葉の論述の品質は言語的情報のみを考慮することで推定することが可能であるが，話し言葉の陳述は言葉から解釈できる情報のみではなく，発言者の動作などの非言語的要素からも影響を受けることが知られている [4]．そのため，複数人議論の自動評価・フィードバックの実現に向けて，複数のモダリティの情報を収録した議論コーパスと，それらの情報を基に複数人による相互作用を考慮する議論の品質評価モデルの構築が必要となる．

このような背景から，本研究では議論の品質評価ラベルが注釈付けされたマルチモーダル複数人議論コーパスの構築，および品質評価モデルの構築と性能評価，

分析を実施する．本論文は本章を含めて 7 章から構成される．2 章では，言語資源，および品質推定についての関連研究をまとめる．3 章では，日本語による対面対話を収録した複数人議論コーパス構築について説明する．4 章では，3 章で構築したコーパスの議論に対して品質スコアを評定する方法について説明する．5 章では，複数人議論の品質を自動評価する手法について提案する．6 章では，5 章の手法を 3 章，4 章で構築したデータセットに対して適用した結果を報告する．最後に，7 章で本論文の内容，および今後の展望・課題をまとめる．

第2章　関連研究

本章では，複数人対話や議論の品質評価に関する関連研究をまとめながら，本研究における問題意識や立ち位置についてまとめる．はじめに，2.1 節に複数人インタラクションを研究するために収録された資源（コーパス）について俯瞰する．その次に，2.2 節では書き言葉・話し言葉の議論を対象とした品質評価に関する研究をまとめる．最後に，2.3 節に近年構築された議論の品質を評価する軸を体系的にまとめた理論について述べる．

2.1　コーパス

複数人によるインタラクションの分析や理解を試みる研究は古くから関心が持たれている．複数人議論を収録し，利用可能なコーパスとして配布されているものの代表に AMI コーパス [5] がある．このコーパスは，タイムスタンプ付きの発話の書き起こしや映像・音声データに加えて，対話の参照要約や部屋に用意されているホワイトボードの書き込みなど対話中の多種多様な情報が付与されている．その他にも，ICSI コーパス [6] や国際電気通信基礎技術研究所の構築したマルチメディアデータベース [7] など数多くのコーパスが存在する．Aran らは複数人の議論場における各議論参加者の支配度や個人特性を注釈付けしたコーパスを構築している [8]．MULTISIMO コーパスは，Family Feud game[1] を用いて，複数人がタスクを遂行するためにお互いがどのような議論や協調を行うかを調査する目的で構築されたコーパスである [9]．これらのコーパスは主に，自然言語処理が対象とする複数人対話の要約や対話の構造解析，インタラクションの研究が対象とする人と人の関係や議論場の状態を理解することを目的として構築されている．

様々なコミュニケーションの研究が 2000 年代前後から活発に実施されているが，対面対話を対象とする議論や議論参加者の陳述の品質を対象とした研究は活発に行われてこなかった．しかし，近年では非定形知の重要性が認識されると共に，議論や議論参加者の品質を対象とする研究を目的としたコーパスが構築されてきている．Zhang らは Oxford 式のディベートをコーパスとして構築している [10]．Oxford

[1] http://familyfeudfriends.arjdesigns.com

式ディベートとは，ある議題に対して 2 つのチームが対立する形を取り，どちらのチームが聴衆をより説得できるか争う競技式ディベートである．勝敗はディベート開始前後に実施される聴衆の投票に基づいて決定される．Zhang らはディベート中に発生する会話の流れに着目した勝者予測モデルを構築し，60%ほどの正解率で予測が可能であることを報告している．ディベートを対象とした言語資源の構築に関する研究としては，若手議員のディベートスキルを訓練するチュータリングシステムの構築を目的としたもの等も存在する [11]．一方で，複数人による合意形成や協調作業に着目したコーパスも構築されている．林らは議論参加者のコミュニケーションスキルを自動推定することを目的として，就職試験を想定したグループディスカッションコーパスを構築している [12]．林らはディスカッションの形式が「自由討論」「インバスケット」「ケーススタディ」「ディベート」の 4 種類に分類可能なことを示し，採用試験で用いられることの多いインバスケット型・ケーススタディ型のディスカッションを収録している．このコーパスには発話の書き起こしや Web カメラなどの機器から得られる動作・音声データ，および「傾聴する姿勢」や「情報伝達力」などコミュニケーションスキルに関連する 5 つの項目について 5 件法で議論参加者を評価した結果を付与している．Olshefski らは教室の協調学習の分析を目的として，英語圏の高校の国語の授業で行われる議論を 29 対話収録したコーパスを構築している [13]．各発話の書き起こしに対して議論構造を分類するタグを付与し，それらを推定するベースラインモデルを構築している．本研究においては，林らや Olshefski らと同様に複数人による対面形式の議論を対象としているが，本研究では議論参加者の陳述内容や，複数人によって実施される議論そのものの品質を評価する目的でコーパスを構築している点で既存研究と異なる．

2.2 品質評価

1 章でも述べた通り，書き言葉の議論を対象とした品質推定は様々行われている．代表的なものとしては英語学習者が語学学習の際に執筆するエッセイの論/議論[2]と品質スコアをペアデータとしたコーパスを構築し，議論の強さ [15] や，構成

[2]ここでは論（argument）と議論（argumentation）の違いについて簡単な例で説明をする．例えば,「小学校からの英語教育を義務化すべきか」といった課題のエッセイが存在した場合に「私は小学校からの英語教育を義務化すべきであると考える．1 つめの理由としては，第二言語を習得することは第二の魂を持つとを言われており，..... 2 つめの理由としては，幼い頃から第二言語を習得しておくことで多様な価値観に触れることが可能となる点が挙げられる．..... 以上のことから，私は小学校からの英語教育を義務化すべきと考えている.」といった形のエッセイがあるとする．このとき,「1 つめの理由と根拠」「2 つめの理由と根拠」といったエッセイを支える一つ一つの要素が論，それらの論が組み合わさってできる 1 つの全体像が議論と定義されている [14].

[16]，容認性 [17] など様々な評価軸によって自動評価する AES がある．これらの品質評価タスクでは主に，入力データと正解スコア（絶対値）のペアを学習したモデルを構築する教師あり学習によるアプローチが主流となっている．しかしながら，学習モデルを構築する際にスコアの品質を確保することが非常に困難であることが知られている [18]．そこで，1 つの論/議論に対して品質ラベルを付与する方法ではなく，2 つの論/議論のペアを作成し，そのどちらがより評価軸に対して高いスコアであるかを付与するペアワイズなアノテーション手法を用いた研究も行われている．Habernal らは Web から抽出した同一トピックに言及している議論ペアのどちらの方が納得感が高いかをクラウドソーシングを用いてアノテーションしたデータを作成し，そのラベルを推定する手法を提案している [19][20]．Habernal らの研究に影響を受け，より効果的なペアワイズ評価を模索する研究も存在し [21]，今後も比較評価による研究は盛んになっていくと推測される．一方，論/議論に対して品質ラベルを付与する作業を必要としない品質推定の研究も存在する．Wachsmuth らは論の主張に用いられている根拠がどの程度主張と関連しているかを正解ラベルを学習せずに算出する手法を提案している [22]．これは「主張を行う際に用いられる根拠が同一の主張を行う人から多く言及されているならば，その根拠は主張に対して関連性がある」という仮説の基，PageRank[23] の原理を議論データに対して応用したものである．以上にまとめたように，議論の品質評価の研究は主に，エッセイや Web テキストから収集した議論をターゲットとして実施されているのが現状である．一方，対面形式の議論の品質評価では言語情報のみならず発言者の動作や音声など様々な情報を考慮しなければいけないため，マルチモーダルな情報を利用した評価モデルの構築が必要となる．

インタラクションを対象とした研究では人やコミュニケーションのパフォーマンス関して評価する手法を提案する研究は数多く存在する．議論参加者のコミュニケーション能力を自動評価する手法を岡田らが提案している [24]．岡田らは人事採用経験者に林らの構築したグループディスカッションコーパス [12] の各議論参加者を「傾聴する姿勢」「双方向の円滑なコミュニケーション」「意見集約力」「情報伝達力」「論理的で明瞭な主張」「総合的なコミュニケーション能力」の軸で評価する作業を依頼し，獲得したスコアを推定するマルチモーダルな評価モデルを提案している．また，設計した各モダリティの特徴量と能力値との関係性について SVM の学習平面を基に分析している．他にも，Dong ら [25] や Avci ら [26] は議論参加者や議論状態から獲得できる情報を用いてグループの性能を自動評価する手法を提案している．インタラクションの研究では主にパラ言語・非言語情報を用いてモデルを構築することが一般的であることに着目し，Murray らはパラ言語・非言語情報に発話から得られる言語的特徴を加えた状態での評価実験を行い，言語情報の有効性を示している [27]．Miura らは任意のタスクを実施するグループの

アウトプットの品質をプロダクトディメンジョンに基づき評価するマルチモーダルな手法を提案している [28].

以上に示したように現在自然言語処理の分野では書き言葉による議論を対象に研究が進められているが，複数人による議論の品質そのものを自動評価する研究は著者の知る限り存在しない．インタラクション分野の研究においても，グループ活動の性能やそれによって生み出された結果，個人の性能を評価する研究は行われているが，議論内容そのものの品質をターゲットとした研究は盛んではない．そこで本研究は，インタラクション研究の知見を取り入れながら，自然言語処理の分野で取り扱うような議論の品質推定タスクを複数人議論のデータを対象に実施する．

2.3 品質評価の理論

2.2 節で紹介したように議論の品質評価の研究は様々行われてる．しかし，それらの研究が推定の対象としている評価軸は各自の研究タスクに最適化されており，任意の議論に対して普遍的な評価軸ではないものも存在する．それらを問題視して，Wachsmuth らは [29] は過去に実施された論（argument）および議論（argumentation）の分析や品質評価の研究を基に，計算論的アプローチに基づく議論の品質評価の理論体系を構築している．本研究のデータセット構築においても Wachsmuth らの分類および定義を用いるため，本節では彼らの理論の概要をまとめる．

Wachsmuth らの定義では論/議論の品質は「適切性（Cogency）」「有効性（Effectiveness）」「合理性（Reasonableness）」の 3 種類の主要評価軸によって評価できるとしている．適切性は主に論を評価するのに適している一方，有効性と合理性は一般に議論を評価するのに利用される．それぞれの主要評価軸は複数の従属評価軸を有している．表 2.1 に論/議論の品質評価の分類体系をまとめる．

適切性（Cogency）の定義は “An argument is cogent if it has acceptable premises that are relevant to its conclusion and that are sufficient to draw the conclusion.” とされており，その論で用いられている前提が結論に対して十分かつ関連しており，一般的に容認される状態であれば適切性が高いと判断される．この適切性に従属する 3 つの軸はローカルな容認性（Local acceptability），ローカルな関連性（Local relevance），ローカルな充足性（Local sufficiency）でそれぞれ以下のような定義となっている．

ローカルな容認性（Local acceptability） A premise of an argument is acceptable if it is rationally worthy of being believed to be true.

表 2.1 論/議論の品質評価の分類体系

評価軸	省略形	和訳
Cogency	Co	適切性
Local acceptability	LA	ローカルな容認性
Local relevance	LR	ローカルな関連性
Local sufficiency	LS	ローカルな充足性
Effectiveness	Ef	有効性
Credibility	Cr	信用性
Emotional appeal	Em	情動性
Clarity	Cl	明瞭性
Appropriateness	Ap	妥当性
Arrangement	Ar	順序性
Reasonableness	Re	合理性
Global acceptability	GA	グローバルな容認性
Global relevance	GR	グローバルな関連性
Global sufficiency	GS	グローバルな充足性

ローカルな関連性（Local relevance）A premise of an argument is relevant if it contributes to the acceptance or rejection of the argument's conclusion.

ローカルな充足性（Local sufficiency）An argument's premises are sufficient if, together, they give enough support to make it rational to draw its conclusion.

端的に述べると，適切性は論に関する論理的な側面を評価する軸となっていて，議論中に存在する論に用いられている根拠が受け入れられ，論の主張に対する関連性が認められ，主張を述べるのに十分であれば適切性が満たされることになる．

有効性（Effectiveness）の定義は “Argumentation is effective if it persuades the target audience of (or corroborates agreement with) the author's stance on the issue.” とされており，その議論が対象となる聴衆に対して論者の議題に対するスタンスを納得させる状態であれば有効性が高いと判断される．この有効性に従属する 5 つの軸は信用性（Credibility），情動性（Emotional appeal），明瞭性（Clarity），妥当性（Appropriateness），順序性（Arrangement）でそれぞれ以下のような定義となっている．

信用性（Credibility）Argumentation creates credibility if it conveys arguments

and similar in a way that makes the author worthy of credence.

情動性（Emotional appeal） Argumentation makes a successful emotional appeal if it creates emotions in a way that makes the target audience more open to the author's arguments.

明瞭性（Clarity） Argumentation has a clear style if it uses correct and widely unambiguous language as well as if it avoids unnecessary complexity and deviation from the issue.

妥当性（Appropriateness） Argumentation has an appropriate style if the used language supports the creation of credibility and emotions as well as if it is proportional to the issue.

順序性（Arrangement） Argumentation is arranged property if it presents the issue, the arguments, and its conclusion in the right order.

端的に述べると，有効性は適切性とは異なり，議論の修辞的な側面を評価をする軸となっている．議論の内容が情に訴えかけるもので信用でき，利用する語や構文構造やわかりやすいものであれば有効性が満たされることになる．

合理性（Reasonableness）の定義は “Argumentation is reasonable if it contributes to the issue's resolution in a sufficient way that is acceptable to the target audience.” とされており，その議論が対象となる聴衆が受け入れられる形で議題の解明に貢献している状態であれば合理性が高いと判断される．この合理性に従属する 3 つの軸はグローバルな容認性（Global acceptability），グローバルな関連性（Global relevance），グローバルな充足性（Global sufficiency）でそれぞれ以下のような定義となっている．

グローバルな容認性（Global acceptability） Argumentation is acceptable if the target audience accepts both the consideration of the stated arguments for the issue and the way they are stated.

グローバルな関連性（Global relevance） Argumentation is relevant if it contributes to the issue's resolution, i.e., if it states arguments or other information that help to arrive at an ultimate conclusion.

グローバルな充足性（Global sufficiency） Argumentation is sufficient if it adequately rebuts those counter-arguments to it that can be anticipated.

端的に述べると，合理性は一つ一つの論を大局的に捉えた際に議題に対して十分関連しており，ある程度予期される批判などについての反論も含みながら容認される形となっていれば合理性が満たされることになる．

第3章　コーパスの構築

2章で述べたように，議論の品質評価に向けた複数人による人対人の対面対話を収録したコーパスは著者の知る限り存在しない．そこで，品質評価に向けた複数人議論コーパスを構築する．はじめに，3.1節でコーパスの収録環境や対話設定についてまとめ，その内容を基に実施した収録について3.2節で説明する．3.3節では，収録した映像・音声データを議論の品質評価モデルで取り扱える形としてコーパス化する手法について述べる．更に，3.4節で構築したコーパスの発話を意味的まとまりに分割するトピックセグメンテーションを実施する．

3.1　収録設定

本研究は複数人議論の品質評価に向けてマルチモーダルコーパスの構築を試みる．データの内容を多様で質の高いものにするため，本研究において評価対象となる「各議論参加者の陳述」や「複数人による議論内容」はなるべく多様なものを収録したい．そこで，4名の話者が討論および合意形成を行う対話設定を構築し，「自身のスタンスがより良いことを相手に納得させる」「答えのない議論で全員が合意できるアイデアを提案する」という異なる性質の立論をマルチモーダルデータとして収録する．

3.1.1　収録環境

対話の収録環境を図3.1に示す．部屋に円形テーブルを設置し，計4名の議論参加者を均等な間隔で配置する．対話の様子を俯瞰して記録するためにテーブルの中心に360度カメラ (Ricoh Theta V[1]) を，360度カメラの真上の天井に頭上カメラ (GoPro HERO5 Session[2]) を設置する．また，360度カメラの予備として部屋の壁際に予備のビデオカメラを設置する．加えて，参加者の表情や視線，上半身の動作情報を記録するために各参加者の正面にも上半身カメラ (GoPro HERO5 Session)

[1]https://theta360.com/ja/about/theta/v.html
[2]https://gopro.com/ja/jp/update/hero5_session

図 3.1 収録環境

を設置する．撮影された映像の例として，図 3.2 に (i) 360 度カメラ (ii) 頭上カメラ (iii) 上半身カメラ の 3 つによって撮影された映像のキャプチャ画像を示す．また，議論参加者の発言をノイズを減らした状態で録音するために，各議論参加者にピンマイクレコーダー（TASCOM DR-10L[3]) を装着する．

3.1.2 収録手続き

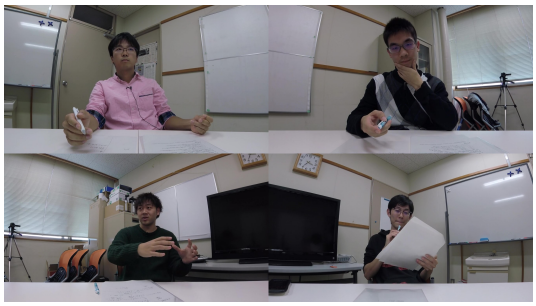
議論の収録手続きは以下の 1 から 3 の要素で構成される．1 対話の収録に必要な時間は合計で 1 時間程度である．本研究で構築するコーパスは 2 および 3 の対話を映像として録画し，構造化したデータを収録している．以下に，各要素の具体的な手続きを示す．

1. **収録に関する説明・サーベイ** (20 分)
   集まった議論参加者 4 名は初めに実験遂行者の指示で 2 名一組のグループに分かれる．その後，実験遂行者から本対話の収録趣旨や議論する命題，自身の所属するグループの立脚点等について 5 分程度説明を受ける．説明終了後，実験遂行者が部屋から退出してから 15 分間，実験遂行者から与えられた資料に基づき実験条件の再確認をする．また，同じタイミングでオンライン上で議論用のサーベイを行い，話し合いのための情報を整理する．資料には「口頭で説明した収録の流れ」「議論における留意点」「同じ立脚点の主張のサンプル」を記載している．

[3]https://tascam.jp/jp/product/dr-10l/top

(i) 360 度カメラ

(ii) 頭上カメラ



(iii) 上半身カメラ

図 3.2 各カメラによる撮影映像のキャプチャ

2. **討論** (20 分)
   議論参加者 4 名は前フェーズで整理した内容を用いて 20 分間グループ対抗の討論を行う．本研究では討論の形式に命題単一型の立脚点混合タイプ [14] を採用する．例えば「小中校の教材はタブレットに置き換えるべきである」という命題に対して「置き換えるべき」という肯定的立脚点と「置き換えるべきでない」という否定的立脚点がそれぞれ主張を展開する討論が命題単一型の立脚点混合タイプである．このとき，各議論参加者は話し合い以外にボールペンで手元の資料にメモを残すことは許されているが，電子端末を使い追加のサーベイを行うこと等は許されていない．この条件は対話とは関係のない動作・音声情報を排除するため，そして議論参加者に会話に集中してもらうために設定した．

3. **合意形成** (20 分)

討論終了後から更に 20 分間を目安に，話し合った内容を基にして各立脚点が納得する合意形成を行う．例えば「小中校の教材はタブレットに置き換えるべきである」という命題に対して肯定的/否定的立脚点の両方が納得のいく折衷案として「導入コストや学習者の集中力欠如を考慮して理科などタブレット導入によって学習効果の向上が期待できる科目については導入する」などを導く議論を行う．討論のフェーズと同様の理由で，合意形成においても各議論参加者は話し合いとボールペンで手元の資料にメモを残すことのみ許可されている．

3.1.3 議論内容

本研究では議論において題材とする命題を procon.org[4] に掲載されている issues を参考に 5 つ作成した．命題のリストを表 3.1 に示す．本研究はディベートポータルで行われるような匿名ユーザによるチャット形式のディベートとは異なり，議論参加者に対面形式で行う議論を依頼するため，コーパスとしての質の担保や倫理的問題を考慮する必要がある．そこで，命題選定の際には「命題は個人の主義/信条による心的負担が少なく，かつ知識差によって一方の立脚点が有利または不利になりにくいものにする」という点を考慮した．例えば,「死刑制度は秩序の担保のために必要である」「任意の理由による中絶を法的に認めるべきである」など，倫理観を取り扱う命題を依頼した場合，議論参加者のバックグラウンドや考え方によっては心的な負担を与える可能性が大きいと考える．また,「イランとの核合意は破棄されるべきである」「日本と韓国の間で締結されている GSOMIA を破棄すべきである」といった命題は特定の議論参加者が専門的知識に長けている場合に，残りの参加者が発言や反論しづらくなり，議論が停滞する可能性があると考える．

3.2 データの収録と整形

前節で設計した収録を実施するために，九州工業大学に所属する計 13 名の学部 4 年生および大学院生（修士課程・博士課程）に対して収録の協力を依頼した．4 名の被験者を 1 つのグループとして，合計 5 グループ 10 対話の収録を行った (一部の学生は 2 つのグループに属している)．収録開始前に各議論参加者は「研究室の研究用途に限定したデータの利用」「個人を特定できる情報を匿名化した対話データの Web 上への無償公開」「研究室の利用許可を得た外部機関へのデータ配布/公

[4]https://www.procon.org/

表 3.1 議論命題一覧

命題 ID	命題
T0	(小中高の) 生徒は制服を着用すべきである
T1	成人の拳銃所持・携帯の権利を認めるべきである
T2	小中高の教材はタブレットに置き換えるべきである
T3	飲酒可能年齢は 20 歳から下げられるべきである
T4	未成年の暴力的ゲームのプレイを禁止すべきである

開」という 3 つのデータの取り扱い方に関して，同意/不同意の意思表示と署名を行った．更に，収録終了後に「討論」「合意形成」「2 種類の議論」の内容に関するアンケートを実施し，結果を収集した．表 3.2 にアンケートの質問と回答形式の一覧を示す．収録実験後，10 対話それぞれについて以下の映像・音声データを獲得した．

- 全議論参加者を同時撮影した 360 度カメラの映像
- 全議論参加者を同時撮影した予備のビデオカメラの映像
- 全議論参加者を同時撮影した頭上カメラの映像
- 各議論参加者を撮影した上半身カメラの映像
- 各議論参加者の発言を収録した音声データ

これらの映像・音声データは現在それぞれ独立した機材で収録した状態にあるため，1 対話のデータとしてまとめるためには時系列方向に対しての同期作業が必要となる．そこで，はじめに 360 度カメラの映像から討論と合意形成の開始時刻と終了時刻を著者が映像を閲覧しながら判定し，対話の映像を手動で切り出した．次に，PluralEyes4[5]を用いて，切り出した 360 度カメラの映像の開始時刻と終了時刻に対応する頭上カメラ，および上半身カメラの映像のフレーム数を獲得した．PluralEyes4 はメディアデータの音声に基づきフレーム単位で任意の数のメディアデータの時刻を同期するツールである．それらのフレーム情報を用いて FFmpeg[6] により頭上カメラ，上半身カメラによって撮影された映像を切り出すことで対話の映像を獲得した．PluralEyes4 は先ほども述べたように音声データを基に同期を

[5]https://www.redgiant.com/products/shooter-pluraleyes/
[6]https://ffmpeg.org/

表 3.2 アンケートの質問一覧

質問文	回答形式
討論は上手くいったと感じますか？	7段階評価
どちらの立脚点の方が主張として説得力があると感じましたか？	7段階評価
議論参加者を説得力の高かった順にランキングしてください	ランキング
自身の立脚点の意見を思い通りに主張できましたか？	7段階評価
討論前の段階であなた自身はどちらの立脚点でしたか？	選択式
討論後のあなた自身の立脚点はどちらですか？	選択式
折衷案の模索は上手くいったと感じますか？	7段階評価
決定した折衷案は何ですか？	記述式
決定した折衷案に納得感はありますか？	7段階評価
決定した折衷案はどちらの立脚点により近いですか？	7段階評価
議論は滞ることなく建設的に進行しましたか？	7段階評価
誰がより議論を深める建設的言動をとっていましたか？	ランキング
誰が最も議論をコントロールしていたと思いますか？	選択式

行うため，360 度カメラの映像音声とピンマイクで収録した音声を同期することも可能である．そこで，同様の手続きで各議論参加者のピンマイクによって録音された音声データも獲得した．

以上の整形作業も含め，本研究では 4 人一組による討論および合意形成対話をそれぞれ 5 対話，計 10 対話分の加工済み映像・音声データを獲得した．加工済データ一覧を以下に示す．

- 全議論参加者を同時撮影した 360 度カメラの映像
- 全議論参加者を同時撮影した頭上カメラの映像
- 各議論参加者を撮影した上半身カメラの映像
- 各議論参加者の発言を収録した音声データ

3.3 各モダリティの情報の構造化

前節までの手続きを通して，それぞれ独立のデバイスで収録した映像・音声データを同期し，10 対話分のデータを獲得した．しかし，映像・音声には個人を特定

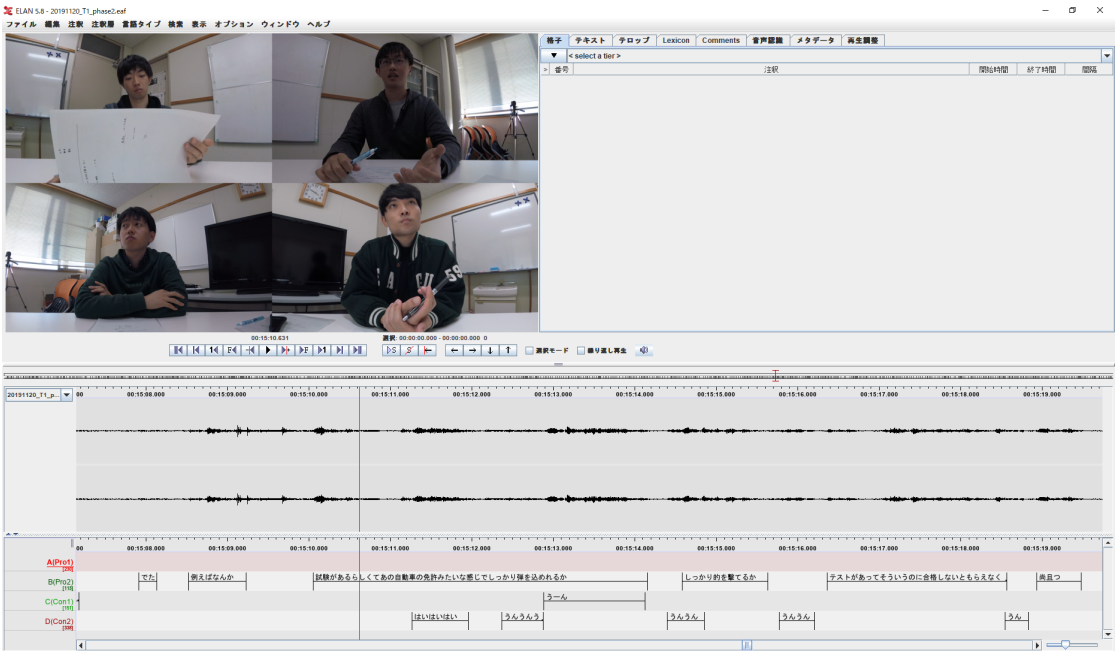


図 3.3 ELAN による書き起こしの環境

できる可能性のある情報が含まれており，それらのデータを含んだまま資源化することは人権に関連する規定等に抵触する恐れがある．加えて，獲得した同期済みの映像・音声データから直接的に議論参加者に関する情報（例えば，発言内容や視線の向き）を獲得することは困難である．そこで，本節では獲得した同期済みの映像・音声データから各議論参加者の言語情報，動作情報，顔情報，音声情報を獲得し，個人の情報が特定できないマルチモーダルコーパスを構築する．

はじめに，言語情報として各議論参加者の発言内容をテキスト化したデータ（以降，発話の書き起こし）を獲得する．上半身カメラで撮影した 4 名分の映像を結合した動画（図 3.2 の (iii)）を作成し，音声動画アノテーションツールである ELAN[7][30] を用いて発話を書き起こした．図 3.3 に ELAN の作業画面を示す．図 3.3 の上段および中段にそれぞれ読み込んだ動画・音声データが可視化され，画面の横方向に時刻が進む UI となっており，ソフト内で動画を再生しながら図 3.3 下部のアノテーション欄に各議論参加者の発言内容を人手で書き起こした．なお，書き起こしの際の発話単位については日本語話し言葉コーパス（CSJ）[31] の物理指標に基づく認定単位に準拠，具体的には 0.2 秒以上の無声区間を転記単位の区切りとした．以上の作業を通じて，各議論参加者の発話について話者 ID，発話開始時刻，発話終

[7]https://archive.mpi.nl/tla/elan

表 3.3 発話の書き起こしデータ例（命題 ID：T2）

話者	開始時刻	終了時刻	発話
B	09:19.475	09:21.272	その始めるタイミング
B	09:21.492	09:24.886	っていうのは必ずしも人は物を扱えるわけではないので
B	09:25.068	09:27.458	どのタイミングから始めても変わらないんじゃないかなと思います
C	09:28.407	09:30.990	小学校 1 年生のときにもうそれをすると
B	09:31.418	09:32.508	それをしてしまっても
C	09:33.003	09:35.759	まだ字も漢字も書けないときにそれをすると
B	09:33.253	09:33.863	えっと

了時刻，発話テキストを 10 対話分獲得した．表 3.3 に発話の書き起こしデータの一部を示す．

次に，動作情報として各議論参加者の上半身の骨格特徴点を獲得する．本研究では，上半身カメラの映像に対して OpenPose[8][32] を利用する．OpenPose とは動画における各フレーム画像から人の骨格，手，顔，そして足の特徴点の座標を検出することの可能なツールである．本研究で獲得した各議論参加者の上半身カメラの映像に対して OpenPose の BODY_25 モデルを適用し，各議論参加者の各フレームにおける骨格，および手の特徴点を抽出した．図 3.4 に本研究のデータに OpenPose を適用した例を示す．

次に，顔情報として各議論参加者の顔特徴点や視線，表情などを獲得する．本研究では，上半身カメラの映像に対して OpenFace[9][33] を利用する．OpenFace とは動画における各フレームの画像から顔特徴点の座標，Facial Action Units[10] (AUs)，視線方向，頭の向き・回転角などを検出することの可能なツールである．本研究で獲得した各議論参加者の上半身カメラの映像に対して OpenFace を適用し，各議論参加者の各フレームにおける顔特徴点の座標，視線方向，頭の向き・回転角，AUs の検出結果を抽出した．図 3.5 に本研究のデータに OpenFace を適用した例を示す．

最後に，音声情報として各議論参加者の声の大きさや高さなどを獲得する．特徴量抽出のための前処理として，獲得した音声データを発話の書き起こしに付与さ

[8]https://github.com/CMU-Perceptual-Computing-Lab/openpose
[9]https://github.com/TadasBaltrusaitis/OpenFace
[10]https://www.cs.cmu.edu/ face/facs.htm

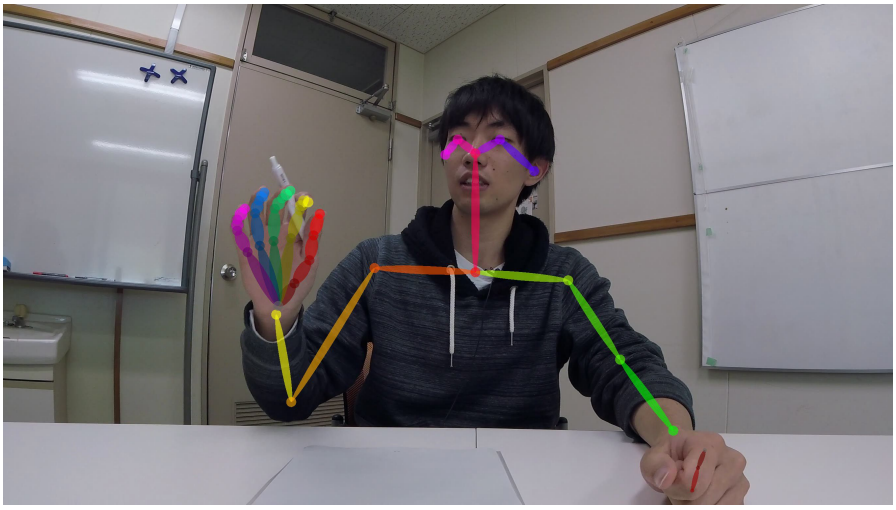


図 3.4 OpenPose による骨格特徴点の検出例

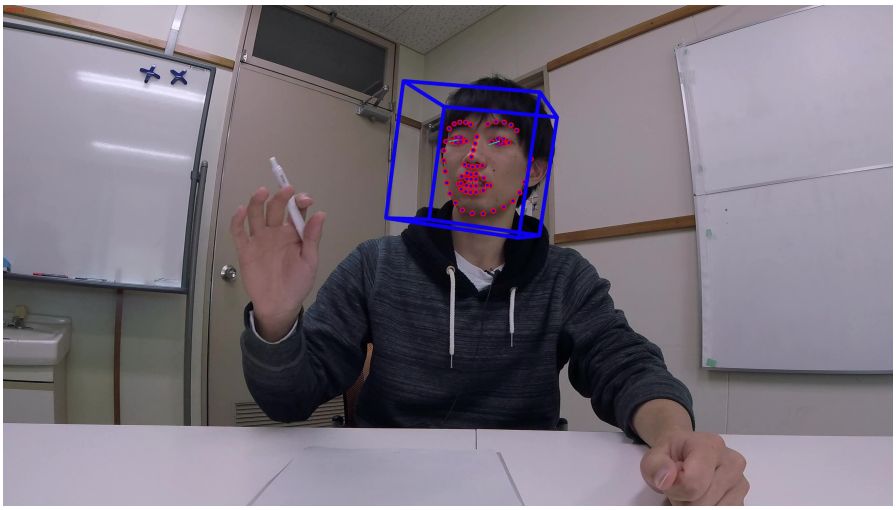


図 3.5 OpenFace による骨格特徴点の検出例

れている発話開始時刻および発話終了時刻を基準にデータ分割を行い，各発話の音声ファイルを獲得した．その後，それらの音声ファイルに対して Surfboard[11][34] を利用した．Surfboard とは音声ファイルから MFCC や RMS 等の音声特徴量を算出することができるオープンソースの Python ライブラリである．本研究で獲得した各議論参加者の発話音声ファイルに対して Surfboard を適用し，各ファイルにお

[11] https://github.com/novoic/surfboard

表 3.4 収録対話の発話数・議論時間の統計値

収録年月日	命題 ID	議論形式	発話数	議論時間
20191114	T2	討論	541	20:00
		合意形成	598	19:59
20191120	T1	討論	832	21:47
		合意形成	796	20:21
20191202	T3	討論	799	20:04
		合意形成	865	20:07
20191205	T0	討論	742	20:07
		合意形成	849	20:16
20191211	T4	討論	615	20:00
		合意形成	812	20:02

ける 13 次元 MFCC，RMS，基本周波数，スペクトル重心，の時間方向の最小値，最大値，平均値，標準偏差，およびジッタとシマの値を抽出した．13 次元 MFCC は発話時の口や喉の形を表現する声道特性，RMS は声の大きさ，基本周波数は声の高さ，スペクトル重心は声の明るさ，ジッタは声の高さのゆれ，シマは声の大きさのゆれとそれぞれ対応している．

以上の手続きをもって，本研究では個人を特定できる情報を排除したマルチモーダル複数人対話コーパスを構築した．以下に，各対話に対して獲得した最終的な構造化データを示す．

1. 発話の書き起こしテキスト（CSV 形式）
2. OpenPose による骨格および指の特徴点（JSON 形式）
3. OpenFace による顔特徴点，視線，頭の向き，および AUs（CSV 形式）
4. 発話音声の特徴量（CSV 形式）

加えて，表 3.4 にコーパスの発話数および議論時間を示す．本研究で構築したコーパスは合計発話数 7,449 発話，合計議論時間約 200 分となった．

3.4 トピックセグメンテーション

一般的に，人の対話中には様々な話題（トピック）が出現する．対話中の発話を意味的なまとまりで分割することは，対話を対象とした研究（例えば，対話要約や談話構造推定）にとって有益とされており，様々な研究が行われている [35][36][37]．本研究が対象とする議論の品質評価においても，意味的に分割されたセグメント単位があれば，それらの分割単位を推定対象としてモデルを構築することが可能になるなど，セグメンテーションを行うメリットが大きいと考えられる．そこで，本研究で構築したコーパスに対してトピックセグメンテーションを実施する．

3.4.1 セグメント境界の定義

山村ら [36] が対象としているコーパスのように，議論参加者が言及するトピックが予め明確であれば，アノテーションするトピックラベルをプリセットとして用意が可能である．しかし，著者が収集している対話データは対話ごとにトピックやドメイン，対話形式が大きく異なるため，アノテーション用のトピックを予め定義することは不可能である．そこで，本研究では対話研究において有名なコーパスの 1 つである AMI Corpus のトピックセグメンテーション [38] を参考にトピックの境界を定義した．先ほど言及した通り，対話中の発話は何かしらのトピックについて言及していると仮定できる．言い換えると，対話の各発話にはメイントピックが含まれており，更に必要に応じてメイントピックを細かくカテゴライズするサブトピックも含まれていると考えられる．その仮定を基に本研究では以下の 4 つをトピックセグメントの境界として定義する．

1. メイントピックが遷移したとき
   例えば,「野球」の話題から「サッカー」の話題へ移った場合.
2. 同じメイントピック内でサブトピックが遷移したとき
   例えば,「野球の NPB」の話題から「野球の MLB」の話題へ移った場合.
3. 話の粒度が細かくなったとき
   例えば,「野球」の話がより詳細になり,「野球の甲子園」の話題へ移った場合.
4. 話の粒度が粗くなったとき
   例えば,「野球の甲子園」の話がより粗くなり,「野球」の話題へ移った場合.

3.4.2 セグメンテーション実験

本研究の発話の書き起こしデータに対し対話のセグメンテーションを実施する．実施にあたって，本研究では 1 対話につき 2 名の作業者を割り当てた．なお，これらの作業者はアノテーション対象となる対話の議論参加者とはならないように考慮した．作業者は 3.4.1 節の定義に従い，発話の書き起こしデータに個人でアノテーション境界を付与した．更に，作業者は各対話セグメントに対してメイントピック，およびサブトピックの内容を表す表現を考え，各セグメントを「[メイントピック]（の [サブトピック]）について」という形式でタイトル付けを行った．ただし，議論時間の確認など議論の進行に関係するセグメントについては「議事進行」，議論とは関係のないセグメントについては「雑談」というタイトルを付与することとした．これらの作業を通して得られる発話の書き起こしデータの例を表 3.5 に示す．

表 3.5 セグメンテーション後の発話の書き起こしデータ例（命題 ID：T0）

話者	開始時刻	終了時刻	トピック	発話
C	00:18.535	00:22.115	s2*	け拳銃って殺傷能力がめちゃくちゃ高いじゃないすか
D	00:22.118	00:23.768	s2	そうまあまあそのための道具だし
C	00:23.546	00:24.016	s2	はい
C	00:25.453	00:25.770	s2	で
C	00:25.960	00:28.000	s3**	他の殺傷能力が高いやつ
C	00:28.319	00:28.693	s3	って
D	00:28.629	00:28.909	s3	うん
			...	
C	00:38.461	00:40.118	s3	まあ色々あると思うんですけど
D	00:39.788	00:40.195	s3	うん
C	00:40.815	00:42.225	s4***	今挙げた車ナイフ
C	00:42.723	00:43.749	s4	ダイナマイトって
C	00:44.079	00:45.596	s4	殺傷能力よりも
C	00:46.867	00:49.877	s4	人間の生活の利便性の方が勝ってるから
C	00:50.422	00:51.819	s4	なんか使用が許可されてる
C	00:52.053	00:52.913	s4	と思うんですよ
D	00:52.346	00:53.219	s4	うんうんうん
C	00:54.110	00:55.650	s5****	でも拳銃の場合って
C	00:57.210	00:59.370	s5	殺傷能力がメインだからなんか
C	00:59.619	01:01.429	s5	生活の利便性っていうのは
C	01:01.723	01:02.883	s5	欠けてると思うんですよね

* 拳銃の殺傷能力について
** 殺傷能力が高い道具について
*** 殺傷能力が高い道具の利便性について
**** 拳銃の利便性欠如について

また，これらの作業によってどの程度信頼性のあるセグメンテーション結果を獲得できたか確認する必要がある．そこで，表 3.6 に作業者が認定したセグメント境界数とそれらの重複数を示す．表 3.6 の結果より，作業者ごとにセグメント境界の数が大幅に異なることはあまりないことがわかった．また，各作業者が独立で認定を行った境界の約半数程度は一致していることが確認できた．境界数が作業

表 3.6 二名の作業者によるセグメント境界数統計

対話 ID	討論			合意形成		
	作業者 1	作業者 2	重複数	作業者 1	作業者 2	重複数
20191114_T2	21	24	16	23	25	12
20191120_T1	31	33	20	67	52	25
20191202_T3	49	42	21	42	44	15
20191205_T0	28	32	18	41	30	15
20191211_T4	45	22	19	45	42	16

者間で大きく離れているものについて著者が確認したところ，メイントピックやサブトピックの設定が自由なことに起因し，作業者がタイトルに汎用的な語を付与するタイプかより詳細な情報を含む語を付与するタイプかによって違いが生じていることを定性的に確認した．また，データを観察すると，一致していない境界は主に「相槌の発話を前後どちらのセグメントに吸収させるかで異なる」などが原因であることを定性的に確認した．そのため，各作業者間で認定境界が大きくことなることはなく，本セグメンテーション作業の結果には一定の信頼性があると著者は結論付けた．

各作業者が独立に作業した結果に一定の信頼性があることを確認したため，最後に，同じ対話のセグメンテーションを実施した 2 名の作業者に各自のデータを議論しながら統合する作業を実施してもらった．つまり，1 対話に 2 名がトピックセグメンテーションを実施し，その結果を 2 名で統合した結果を最終的なセグメンテーション結果とした．表 3.7 に各対話における最終的なセグメント数を示す．

本アノテーションを通じて，10 対話で合計 378 セグメントを獲得した．1 セグメントあたりの平均発話数は約 20 発話である．

表 3.7 統合後のトピック数

収録年月日	命題 ID	議論形式	発話数	セグメント数
20191114	T2	討論	541	24
		合意形成	598	24
20191120	T1	討論	832	35
		合意形成	796	62
20191202	T3	討論	799	43
		合意形成	865	47
20191205	T0	討論	742	30
		合意形成	849	39
20191211	T4	討論	615	37
		合意形成	812	37

第4章　議論品質の評定

3章において討論と合意形成という性質の異なる対話を計10対話収録し，マルチモーダル複数人議論コーパスを構築した．本研究における最終目標は議論の品質を自動的に評価するモデルの構築，およびそれに必要なコーパスの構築である．そのため，本研究のコーパスに収録されている議論の品質を評定する必要がある．しかし，2章でも述べたように，話し言葉の議論を対象とした品質評価に取り組む研究は著者の知る限り多くない．そこで，本章では，書き言葉をターゲットとして構築された評定手法を改良し，本研究に適用する方法を2種類提案する．4.1節では，各議論参加者の対話中の陳述を評定する手法について，4.2節では，各議論セグメントの品質を評定する手法について説明する．

4.1　議論参加者の陳述の評定

複数人議論における品質の評定の1つのアプローチとして，各議論参加者の陳述内容を評定することが考えられる．もし，各議論参加者の陳述を自動評価する手法が実現できれば，大学入試や就職試験など様々な場面において各個人の能力を比較可能な形で推定可能になるなど，いくつかの利点が考えられる．そこで，英語エッセイの各文章に対して説得力などのスコアを付与している先行研究 [39] を参考に評価軸および評価基準を作成し，評定実験を実施する．

4.1.1　陳述評定スキーム

陳述の評定に用いる評価軸の一覧を表4.1に示す．説得力は議論参加者の陳述に対してどの程度説得力があったかを示すスコアである．納得する人が多いと考えられる陳述であるほど，このスコアは高くなる．発想力は議論参加者の陳述にどの程度具体性や独創性があったかを示すスコアである．つまり，この評価軸は議論参加者の命題に対する考えや知識の深さ，自信などを表したものとなっている．雄弁さは議論参加者の陳述がどの程度雄弁であったかを示すスコアである．自身の考えや意見を流暢にかつ明確に表現できていればいるほどこの評価軸のスコア

表 4.1 陳述を評定する軸

評定軸	スコアレンジ	概要
説得力	1-6	陳述にどの程度説得力があったか
発想力	1-6	陳述がどの程度具体的で独創的だったか
雄弁さ	1-6	陳述がどの程度雄弁だったか
一貫性	1-6	陳述にどの程度一貫性があったか
批判的思考力	1-6	陳述がどの程度批判的であったか
説得戦略	Yes/No	Ethos/Logos/Pathos を有していたか

は高くなる．一貫性は議論参加者の陳述にどの程度一貫性があったかを示すスコアである．つまり，この評価軸は議論参加者の主張が何回も変わることなく一貫した流れになっているかを表している．批判的思考力は議論参加者の陳述がどの程度批判的であったかを示すスコアである．人から与えられる情報や自身の意見/考えを冷静に分析し，客観的であればあるほどこの評価軸のスコアは高くなる．説得戦略は議論参加者の陳述がアリストテレス [40] によって定義される Ethos (話者の信憑性)，Logos (論理性)，Pathos (感情的アピール) を感じさせるものかを示す指標である．各評価軸の基準については付録 A に記載する．

4.1.2 陳述評定実験

各グループに対して第三者 3 名を割り当て，議論参加者の陳述の評定を依頼した．議論の経過に伴うスコアの変化を分析するために，評定者は図 3.2 の (iii) の上半身カメラによる映像を結合した動画を視聴しながら 4.1.1 節で定義した評価指標を基に，議論開始から 5 分おきにその時点までの話者の陳述を評定した．

評定結果について一致率を求め，6 つの評定軸に付与されたスコアの信頼性を確認した．本研究では一致率の算出にクリッペンドルフの α 係数を採用した．この係数は -1 から 1 の連続値で全ての尺度の一致率を算出できるもので，1 に近ければ近いほど高い一致率を示す．経験的に 0.667 以上であればアノテータによる評価が一致しているとされている．表 4.2 に各評価軸のクリッペンドルフの α 係数を示す．表中の全ての値が基準値を下回っており，本研究で実施した時間区切りで議論参加者の陳述を評定する方法では信頼性のスコアが得られなかった．また，20 分時点での各評価軸の一致率を比較すると「発想力」「雄弁さ」「Ethos」「Pathos」の 4 つの評価軸の一致率が特に低い傾向にあった．これは 4 つの評価軸は他と比較してアノテータの感じ方などの主観に特に大きく関わることが原因であると考

表 4.2 各評定軸のクリッペンドルフ α 係数

評価軸	5 分時点	10 分時点	15 分時点	20 分時点
説得力	**0.322**	0.168	-0.023	0.144
発想力	**0.284**	0.035	-0.133	-0.059
雄弁さ	**0.352**	-0.019	0.055	0.072
一貫性	**0.450**	-0.031	0.084	0.167
批判的思考力	0.167	**0.344**	0.198	0.303
説得戦略 (Ethos)	-0.086	**0.221**	0.091	0.068
説得戦略 (Logos)	0.136	0.042	0.148	**0.153**
説得戦略 (Pathos)	-0.111	**-0.053**	-0.147	-0.068

えられる．表 4.2 より,「Logos」を除いた評価軸について一致率が最大となっているのは 5 分時点または 10 分時点となっていることが確認された．つまり，時間が経つほど評価する議論が煩雑になり，スコアの一致率が低くなる傾向があると推測される．今後は単純な時間区切りではなく，議論内容を崩さない細かい分割単位であるトピックや議論構造を用いた陳述の品質の評定を実施，その結果をボトムアップに統合することで一致率の高い評定結果が獲得できると考えられる．

本評定を実施する際に，評定者にはスコアのみでなく，そのスコアに決定した理由について 100 字以内で記述するように求めた．そこで，本節では説得力が低いと評価された話者と高いと評価された話者の間にどのような違いがあったのかを定性的に分析した結果を報告する．定性的分析のために議論参加者を説得力の低い話者 (低群) と高い話者 (高群) を仕分ける．はじめに，2 名以上の評定者が同じスコアを評定した場合にはそのスコアを，3 名の評定者がそれぞれ違うスコアを評定した場合には平均値をその議論参加者の説得力とする集計処理を行った．集計した説得力スコアを基に説得力スコアが 1, 2, 3 の議論参加者を低群，4, 5, 6 の議論参加者を高群として定義する．低群に分類された議論参加者には主に「データや自身の経験など具体的な事例を通して話をしているが，主張したいポイントが聞き手に伝わらない」といった傾向があった．それと同時に「主張を行った後の他人からの批判や他人の主張によって主張が揺らぐ」「相手の陳述に対して言及するに留まり，自身の主張を行わない」といった基準で減点される傾向があることもわかった．一方，高群に分類された議論参加者には「言いたいことを明確に表現できている」といった一般的な傾向の他に「他の議論参加者の陳述に対して言及を行い，同時に自身の立場を明確化する」といった行動が見られることを確認した．この結果から，インタラクションにおける議論においては，主張のみに

よって議論参加者の説得力は同定されないことを定性的に確認した．つまり，インタラクションにおける陳述の評定には話者自身の主張内容のみならず，その話者が他の話者に対してどのような発言や応答を行ったかも大きく影響することを示唆している．

4.2 議論セグメントの品質の評定

4.1 節では各議論参加者の陳述を 6 つの観点から評定する実験を実施した．実験を通して，本研究では 4.1 節に基づく品質の評定手法の問題点を 2 点明らかにした．1 点目は，採点対象となる議論や陳述は物理指標に基づいて分割されていると評定の難易度が上がる点である．2 点目は，単純に評価の対象となる議論参加者の情報のみを取り扱うだけでは不十分，つまり対象となる議論参加者とそれ以外の議論参加者のやりとりも考慮しなければならない点である．これらに加え，現在のデータ構築手法では 1 名の議論参加者にスコアが付与される形となっており，自動評価モデルを教師あり手法で構築する場合に大量の対話データが必要となってしまう．これらの問題を回避する品質の評定方法として，各対話セグメントを 1 つの小さな議論と見なし，その内容を評定することが考えられる．そのアプローチの場合，複数名によって構成される議論の評定値を算出でき，更に各セグメントのスコアを，例えばセグメント内の発話数の割合などで分配すれば各議論参加者の評価も可能となる．そこで，自然言語による議論の計算論的品質評価に関する理論 [29] を基に評定軸および基準を作成し，実験を実施する．

4.2.1 品質評定スキーム

Wachsmuth ら [29] の定義する分類体系では，議論の品質は「有効性（Effectiveness）」「合理性（Reasonableness）」の 2 種類の主要評価軸とそれらに付随する従属評価軸で評価することができる（2.3 節参照）．先行研究を基に，本コーパスで議論セグメントの品質の評定に用いる軸一覧，および各軸で測られるポイントを表 4.3 に示す．実際に評定者に提示した各軸に対する説明については付録 B に記載する．

4.2.2 品質評定実験

陳述評定実験と同様に，各グループに対して第三者 3 名を割り当て，議論参加者の陳述の評定を依頼した．評定者は図 3.2 の (iii) の上半身カメラによる映像を結合

表 4.3 議論セグメントの品質の評定軸一覧

評定軸	略称	スコアレンジ	概要
合理性	Re	(VL, L, M, H, VH)	議論が GA から GS を満たしているか
容認性	GA	(L, M, H)	議論の内容/進行方法が適切で許容できるか
関連性	GR	(L, M, H)	議論の内容が議題と関連したものであるか
充足性	GS	(L, M, H)	議論の内容が内省されているか
有効性	Ef	(VL, L, M, H, VH)	議論が Cr から Ar を満たしているか
信用性	Cr	(L, M, H)	議論の内容を信頼することができるか
情動性	Em	(L, M, H)	議論に対してオープンマインドになれるか
明瞭性	Cl	(L, M, H)	議論の内容が明快か
妥当性	Ap	(L, M, H)	議論における各話者の言動が建設的か
順序性	Ar	(L, M, H)	論点や根拠，結論の流れが理解しやすいか

した動画，およびセグメンテーション済みの発話の書き起こしデータを閲覧しながら 4.2.1 節で定義した評価指標を基に議論セグメントを評定した．ここでの議論セグメントとは，3.4.1 節で定義した境界の内,「メイントピックが遷移したとき」として付与された境界によって分割されたセグメントとしている．理由としては，定義の 2 から 4 の箇所もセグメント境界として評定を実施した場合，発話の意味的な過分割が発生し，4.2 節で述べた問題点が発生してしまうためである．作業手順としてははじめに，それぞれの従属評価軸を L（Low），M（Middle），H（High）で評価し，その後従属評価軸のスコア分布を基に主要評価軸のスコアを VL（Very Low），L（Low），M（Middle），H（High），VH（Very High）の中から決定する．主要評価軸のスコアを決定するルールについては付録 B に記載する．

評定結果について一致率を求め，主要評価軸および従属評価軸の結果の信頼性を確認した．表 4.4 に各軸のクリッペンドルフの α 係数を示す．本研究による実験では一つの基準である 0.667 に到達する指標はなく，提案スキームでは十分な信頼性を有した評定結果を得ることはできないことを確認した．書き言葉のテキストに対して同様の理論をベースにした評定を専門家とクラウドワーカーに依頼し，分析結果を報告している論文が存在する [18]．この論文では同様の指標を用いて「3 名の専門家の評定結果間の一致率」「2 組のクラウドワーカーグループ（5 名）による評定結果の平均値の間の一致率」「クラウドワーカーグループによる評定結果の平均値と 3 名の専門家の評定結果の平均値の間の一致率」が報告されている．

表 4.4 各評定軸のクリッペンドルフ α 係数

Re	GA	GR	GS	Ef	Cr	Em	Cl	Ap	Ar
0.151	0.087	0.029	0.128	0.135	0.032	0.038	0.017	0.076	0.155

それぞれのクリッペンドルフ α 係数は 0.27〜0.51，-0.27〜0.53，-0.17〜0.54 となっており，専門家間でも 0.667 という 1 つの閾値には到達していない．以上のことより，今後はこれらの理論に基づいてより信頼性の高い評定結果を獲得するアノテーション手法の開発が必要であると考えられる．また，先行研究ではクリッペンドルフ α 係数が 0 を下回ることもあり，偶然よりも高い確率でアノテータ間での不一致が発生している．一方で，本研究の評定結果はそれぞれの評価軸のクリッペンドルフの α 係数が 0 を下回っている軸は存在しないため，アノテータ間で偶然の確率を超えた不一致が発生していないことが保証されている．

4.2.3 評定結果の統合

議論セグメントの品質の評定結果に偶然の確率を超えた不一致の発生がないことから，本研究ではこれらのデータを品質評価モデルの構築に用いる．現在，1 つの議論セグメントに対して 3 名の評価者が導き出したスコアが存在している．そこで，2 名以上が同じスコアを付与した場合はそのスコアを，3 名がそれぞれ異なるスコアを付与した場合は平均値を代表スコアとする処理を行い，各セグメントに対する最終的な品質スコアを決定した．表 4.5，図 4.1，および図 4.2 に各評価軸の統合した後の統計値を示す．主要評価軸について観察すると，合理性（Re）およびの有効性（Ef）共に VL および L のラベル数が極端に少ないことが確認できる．このことから，本コーパスに収録されている議論は低品質な議論が少ないと結論づけられる．また，従属評価軸について観察すると，関連性（GR），明瞭性（Cl），妥当性（Ap）は最頻値が H となっていることを確認した．これは今回実施したデータ収録の協力者が学部 4 年生および大学院生（修士課程・博士課程）であり，研究活動などを通じて論理的思考や議論に慣れ親しんでいることに起因すると考えられる．一方で，本研究の最終目標である複数人議論の品質評価モデル作成のためには各スコアのバランスが保たれたデータを構築していく必要がある．そのため今後は，被験者の多様性を増やす，あるいは収録開始前に行うサーベイを実施しない状態での議論を収録するなどの工夫を行い，低品質な議論の収録を目指すことが 1 つの課題としてあげられる．

表 4.5 各評定軸の統合スコア合計値

評定軸	略称	VL	L	M	H	VH	評定軸	略称	VL	L	M	H	VH
合理性	Re	1	12	89	59	17	有効性	Ef	0	9	97	39	33
容認性	GA	-	6	103	69	-	信用性	Cr	-	8	127	43	-
関連性	GR	-	12	73	93	-	情動性	Em	-	9	94	75	-
充足性	GS	-	23	112	43	-	明瞭性	Cl	-	11	81	86	-
-	-	-	-	-	-	-	妥当性	Ap	-	10	60	108	-
-	-	-	-	-	-	-	順序性	Ar	-	19	126	33	-

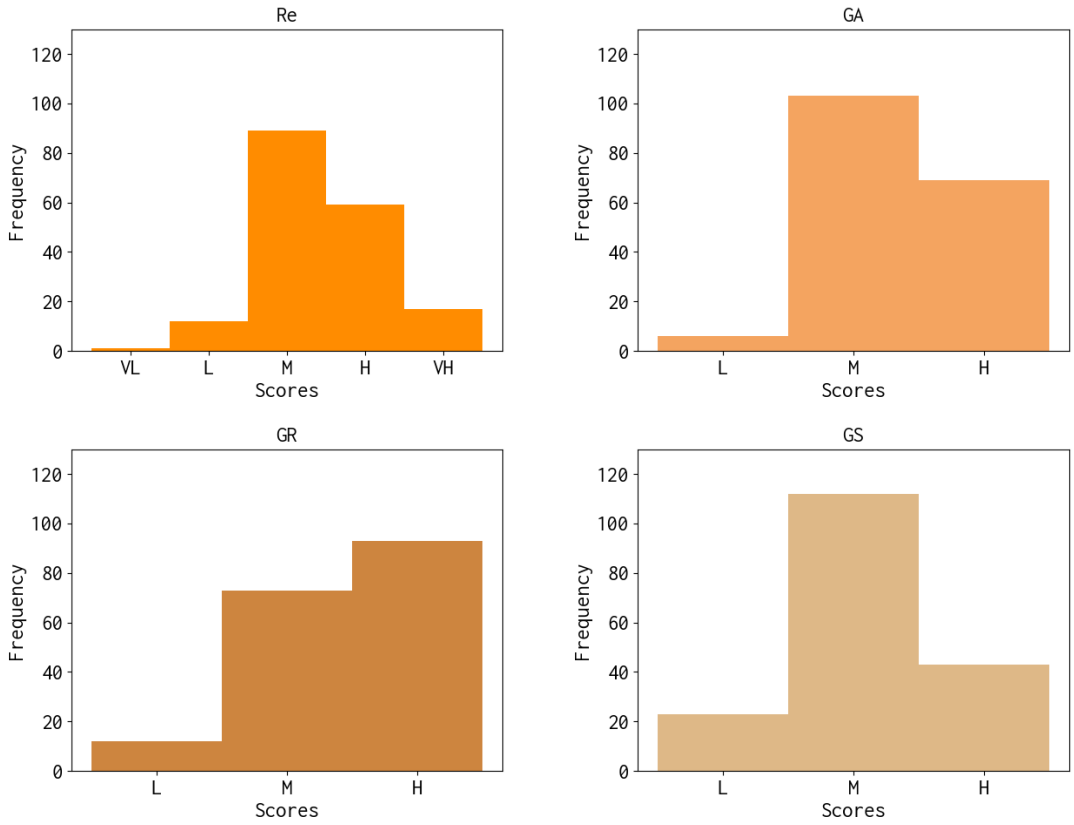


図 4.1 合理性に関する評定軸の統合スコアの度数分布

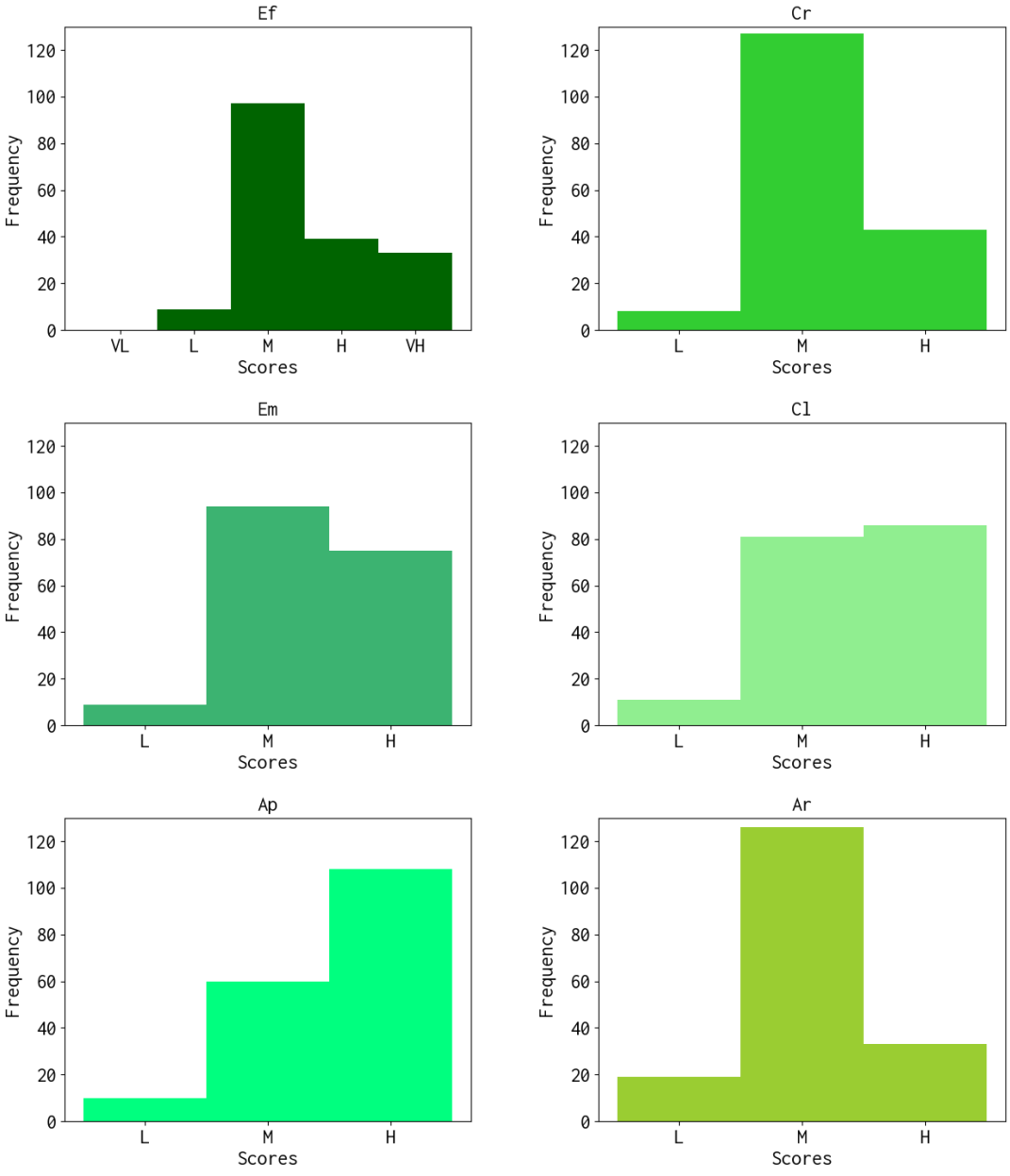


図 4.2 有効性に関する評定軸の統合スコアの度数分布

第5章　品質評価モデルの構築

3章でマルチモーダルな複数人議論コーパスを構築し，4章で議論セグメント1つを10種類の評定軸による観点から評定した．本研究における最終目標は議論の品質を自動的に評価するために必要な資源，および評価モデルの構築である．そこで本章では，構築したデータセットを対象に議論セグメントの品質を自動評価するモデルについて解説する．はじめに5.1節でデータセットや本研究における品質評価タスクを定式化する．次に，5.2節および5.4節にかけて計5種類のモデルを提案する．

5.1　品質評価タスクの定式化

ある議論セグメントSに対して合理性，および有効性に関連する合計10種類の品質ラベルが付与されている．議論セグメントSは発話ベクトル系列$U = \{\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_N\}$を有している．ここで，$N$は議論セグメント$S$に属する発話の数であり，$\boldsymbol{u}_i$は議論セグメント$S$の$i$番目の発話ベクトルである．発話ベクトル$\boldsymbol{u}_i$は以下の形式で表現される．

$$\boldsymbol{u}_i = [\boldsymbol{sp}_i; \boldsymbol{t}_i; \boldsymbol{b}_i; \boldsymbol{f}_i; \boldsymbol{a}_i] \tag{5.1}$$

$\boldsymbol{sp}_i$は議論セグメントSのi番目の発話の話者が$i-1$番目の発話の話者と異なるか否かを表す数，$\boldsymbol{t}_i$はSのi番目の発話のテキスト情報を表現するベクトル，$\boldsymbol{b}_i$はSのi番目の発話中の動作情報を表現するベクトル，$\boldsymbol{f}_i$はSのi番目の発話中の顔情報を表現するベクトル，$\boldsymbol{a}_i$はSのi番目の発話の音声情報を表現するベクトルである．また，$[\cdot;\cdot]$はベクトルの連結を表す．つまり，品質評価タスクは議論セグメントの言語・非言語情報を有する発話ベクトル系列Uを入力として受け取り，指定の軸のスコアy_{dim}を推定するタスクとなる．

5.2　サポートベクターマシン

最もシンプルな品質評価モデルとして，サポートベクターマシン（SVM）[41]を用いたモデルを構築する．図5.1にSVMを用いた品質評価モデルを示す．SVMで

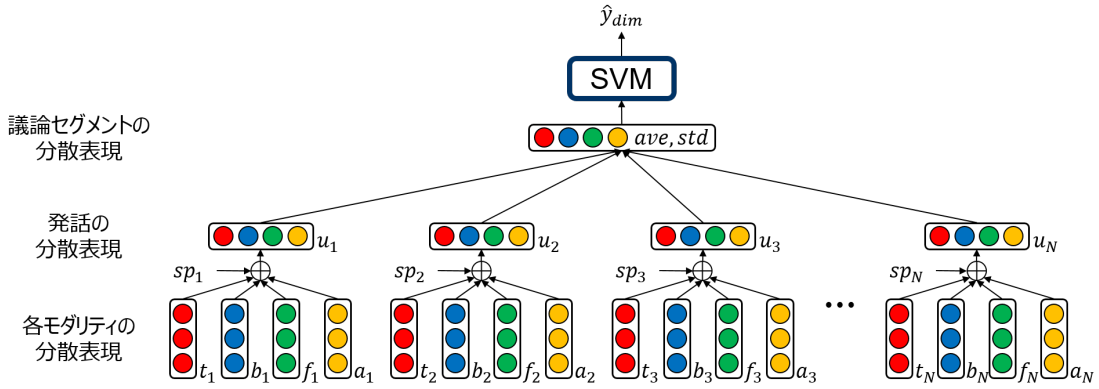


図 5.1 SVM を用いた品質評価モデル

はシーケンスデータを入力することはできないため，本研究では発話ベクトルの各要素の時間方向の平均値および標準偏差を算出し，それを議論セグメントのベクトルとする．そのベクトルを入力として，対象となる評価軸の品質スコア $\hat{y}_{dim}$ を出力する．

5.3 回帰型ニューラルネットワーク

SVM を用いた品質評価モデルでは，各発話ベクトルの各要素の平均値および標準偏差を算出することで議論セグメントベクトルを獲得していた．しかし，この手法では各発話の系列情報を十分に捉えることができない可能性がある．そこで，本節では回帰型ニューラルネットワーク（RNNs）の中で代表的な Long Short-Term Memory（LSTM）[42] と注意機構 [43] を用いた品質評価モデルを提案する．

5.3.1 Long Short-Term Memory

RNNs は系列データを扱えるが，時間方向に遠く離れた過去のデータを任意の時刻 i に反映できないこと（勾配消失問題）が知られている．LSTM は，各中間層のユニットに内部状態を記憶するモジュールである「メモリセル」，次時刻にどの程度メモリセルの情報を保持するか調節する「忘却ゲート」，メモリセルにユニットへの入力値をどの程度反映させるかを調節する「入力ゲート」，そしてメモリセルが次の層にどの程度影響を与えるべきか調節する「出力ゲート」という 4 つの機能により勾配消失問題を解決した RNN モデルである．図 5.2 に LSTM を用いた品質評価モデルを示す．

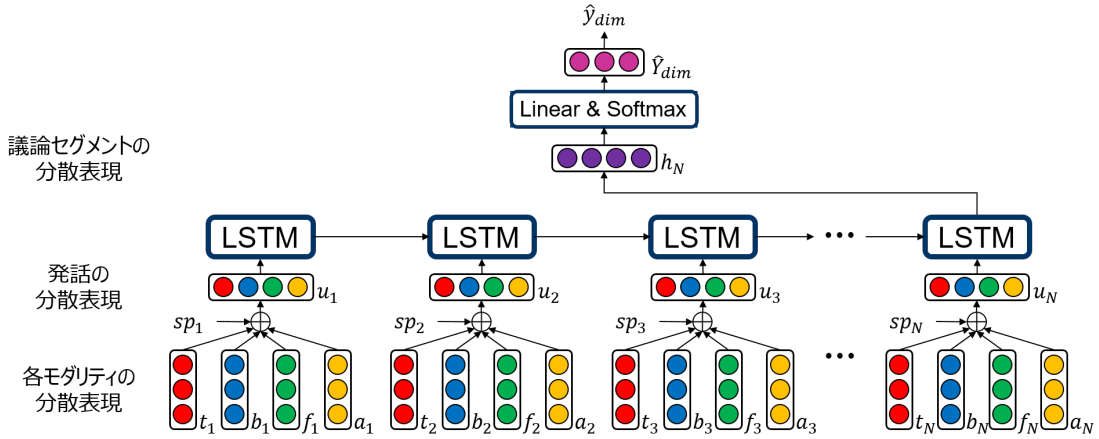


図 5.2 LSTM を用いた品質評価モデル

入力 $\boldsymbol{u}_i$ を受け取った時刻 i における LSTM のユニットは以下のように計算される．

$$\boldsymbol{g}_i^I = \sigma(\boldsymbol{W}_u^I \boldsymbol{u}_i + \boldsymbol{W}_h^I \boldsymbol{h}_{i-1} + \boldsymbol{W}_c^I \boldsymbol{c}_{i-1} + \boldsymbol{b}^I) \tag{5.2}$$

$$\boldsymbol{g}_i^F = \sigma(\boldsymbol{W}_u^F \boldsymbol{u}_i + \boldsymbol{W}_h^F \boldsymbol{h}_{i-1} + \boldsymbol{W}_c^F \boldsymbol{c}_{i-1} + \boldsymbol{b}^F) \tag{5.3}$$

$$\boldsymbol{c}_i = \boldsymbol{g}_i^F \odot \boldsymbol{c}_{i-1} + \boldsymbol{g}_i^I \odot \tanh(\boldsymbol{W}_u \boldsymbol{u}_i + \boldsymbol{W}_h \boldsymbol{h}_{i-1} + \boldsymbol{W}_c \boldsymbol{c}_{i-1} + \boldsymbol{b}) \tag{5.4}$$

$$\boldsymbol{g}_i^O = \sigma(\boldsymbol{W}_u^O \boldsymbol{u}_i + \boldsymbol{W}_h^O \boldsymbol{h}_{i-1} + \boldsymbol{W}_c^O \boldsymbol{c}_i + \boldsymbol{b}^O) \tag{5.5}$$

$$\boldsymbol{h}_i = \boldsymbol{g}_i^O \tanh(\boldsymbol{c}_i) \tag{5.6}$$

$\boldsymbol{g}_i^I$ は時刻 i における入力ゲート，$\boldsymbol{g}_i^F$ は時刻 i における忘却ゲート，$\boldsymbol{g}_i^O$ は時刻 i における出力ゲート，$\boldsymbol{c}_i$ は時刻 i におけるメモリセル，$\boldsymbol{h}_i$ は時刻 i における LSTM の隠れ層，$\boldsymbol{W}_u^I$，$\boldsymbol{W}_h^I$，$\boldsymbol{W}_c^I$，$\boldsymbol{b}^I$，$\boldsymbol{W}_u^F$，$\boldsymbol{W}_h^F$，$\boldsymbol{W}_c^F$，$\boldsymbol{b}^F$，$\boldsymbol{W}_u$，$\boldsymbol{W}_h$，$\boldsymbol{W}_c$，$\boldsymbol{b}$，$\boldsymbol{W}_u^O$，$\boldsymbol{W}_h^O$，$\boldsymbol{W}_c^O$，$\boldsymbol{b}^O$ は学習パラメータである．$\sigma()$ と $\tanh()$ はそれぞれシグモイド関数と双曲線正接関数，$\odot$ はアダマール積を表す．

上記の手順を繰り返すことによって，議論セグメントに属する全ての発話ベクトルが入力された LSTM の最終状態 $\boldsymbol{h}_N$ が獲得できる．本研究ではこの $\boldsymbol{h}_N$ を議論セグメントの分散表現と見なし，出力ラベルの予測確率分布 $\hat{Y}_{dim}$ を求める．

$$\hat{Y}_{dim} = \mathrm{softmax}(\boldsymbol{W}_s \boldsymbol{h}_N + \boldsymbol{b}_s) \tag{5.7}$$

$\boldsymbol{W}_s$，$\boldsymbol{b}_s$ は学習パラメータ，$\mathrm{softmax}()$ はソフトマックス関数を表す．最後に，分布の最大値と対応するインデックスを獲得することで予測品質スコア $\hat{y}_{dim}$ を獲得

する．

$$\hat{y}_{dim} = \arg\max_{y_{dim}} \hat{Y}_{dim} \tag{5.8}$$

次節以降，式 (5.2) から式 (5.6) の記述を簡略化するため，$\boldsymbol{h}_i = LSTM(\boldsymbol{u}_i, \boldsymbol{h}_{i-1}, \boldsymbol{c}_{i-1})$ と表記する．

5.3.2 注意機構付き Long Short-Term Memory

LSTM の導入によって発話の時系列方向の情報を捉えられるようになった．しかし，議論セグメント中の発話には議論の内容に対してあまり重要ではない発話，例えば相槌，などが含まれている可能性がある．これらの情報を議論セグメントの内容を代表する発話と同等の情報として取り扱うと，相槌の情報がノイズとなってしまい評価精度に影響を及ぼす可能性がある．そこで，LSTM に注意機構を導入した先行研究 [44][45] を参考に，注意機構付き LSTM による品質評価モデルを 2 種類提案する．

図 5.3 に注意機構によって重み付けされた LSTM の出力によって議論セグメントの品質を評価するモデルを示す．LSTM を用いた品質評価モデルと同様に，以下の式によって時刻 i の LSTM の状態を求めることができる．

$$\boldsymbol{h}_i = LSTM(\boldsymbol{u}_i, \boldsymbol{h}_{i-1}, \boldsymbol{c}_{i-1}) \tag{5.9}$$

LSTM の各時刻における出力 h_i に対する重み a_i を次式を用いて算出する．

$$m_i = \boldsymbol{\omega}^T \tanh(\boldsymbol{h}_i) \tag{5.10}$$

$$a_i = \frac{\exp(m_i)}{\sum_{j=1}^{N} \exp(m_j)} \tag{5.11}$$

$\boldsymbol{\omega}^T$ は学習パラメータ，exp() は指数関数である．その後，a_i によって重みづけされた各時刻における LSTM の隠れ層 $\boldsymbol{h}_i$ の和を用いて，最終的な状態 $\boldsymbol{h}^*$ を計算する．

$$\boldsymbol{r} = \sum_{i=1}^{N} a_i \boldsymbol{h}_i \tag{5.12}$$

$$\boldsymbol{h}^* = \tanh(\boldsymbol{r}) \tag{5.13}$$

そして，この $\boldsymbol{h}^*$ を議論セグメントの分散表現と見なし，出力ラベルの予測確率分布 $\hat{Y}_{dim}$ を求める．

$$\hat{Y}_{dim} = \mathrm{softmax}(\boldsymbol{W}_s \boldsymbol{h}^* + \boldsymbol{b}_s) \tag{5.14}$$

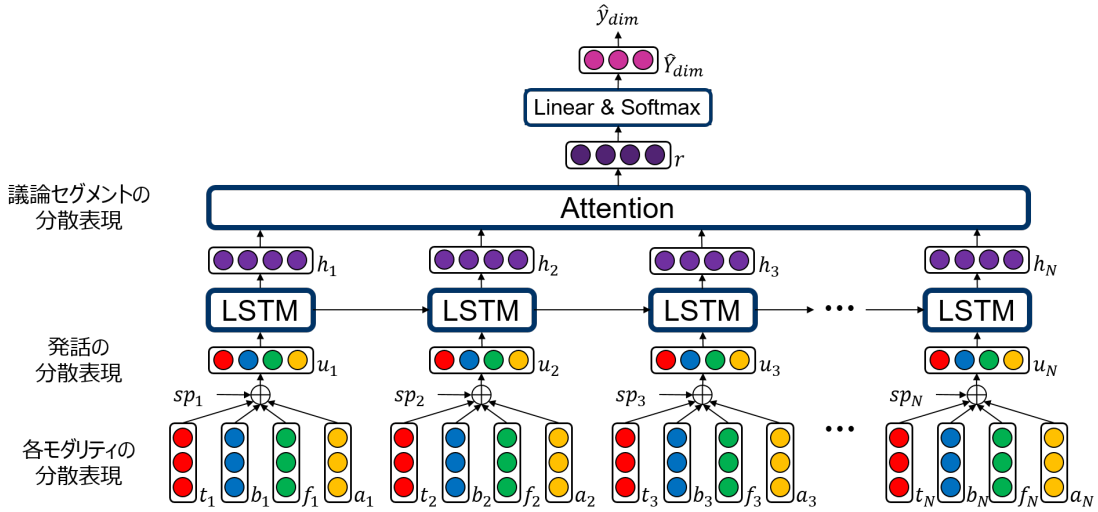


図 5.3 注意機構付き LSTM を用いた品質評価モデル 1

最後に，分布の最大値と対応するインデックスを獲得することで予測品質スコア $\hat{y}_{dim}$ を獲得する．

$$\hat{y}_{dim} = \arg\max_{y_{dim}} \hat{Y}_{dim} \tag{5.15}$$

注意機構の出力のみを用いて品質スコアを予測すると，ノイズを軽減できる期待がある一方，推論に必要な情報を欠落してしまう可能性もある．そこで，図 5.4 のように注意機構による出力と LSTM の最終状態を同時に考慮するモデルも構築する．形式的には，式（5.13）で表現される最終状態を以下のように，注意機構による各出力層の重み付き和 $\boldsymbol{r}$ と LSTM の最終状態 $\boldsymbol{h}_N$ を同時に用いることで実現される．

$$\boldsymbol{h}^* = \tanh(\boldsymbol{W}_r \boldsymbol{r} + \boldsymbol{W}_{h_N} \boldsymbol{h}_N) \tag{5.16}$$

$\boldsymbol{W}_r, \boldsymbol{W}_{h_N}$ はそれぞれ学習パラメータである．後は同様に，最終状態 $\boldsymbol{h}^*$ を用いて予測品質スコア $\hat{y}_{dim}$ を獲得する．

5.4 階層的回帰型ニューラルネットワーク

LSTM や注意機構の導入によって，議論セグメントの発話系列の状態を考慮した品質評価モデルが構築できた．しかし，現在の発話ベクトルで用いられている t_i は単語の系列を直接的に 1 つのベクトルとして表現している問題がある．例えば，話し言葉で「これは絶対に正しい」と「これは正しい，絶対に」といった表現が

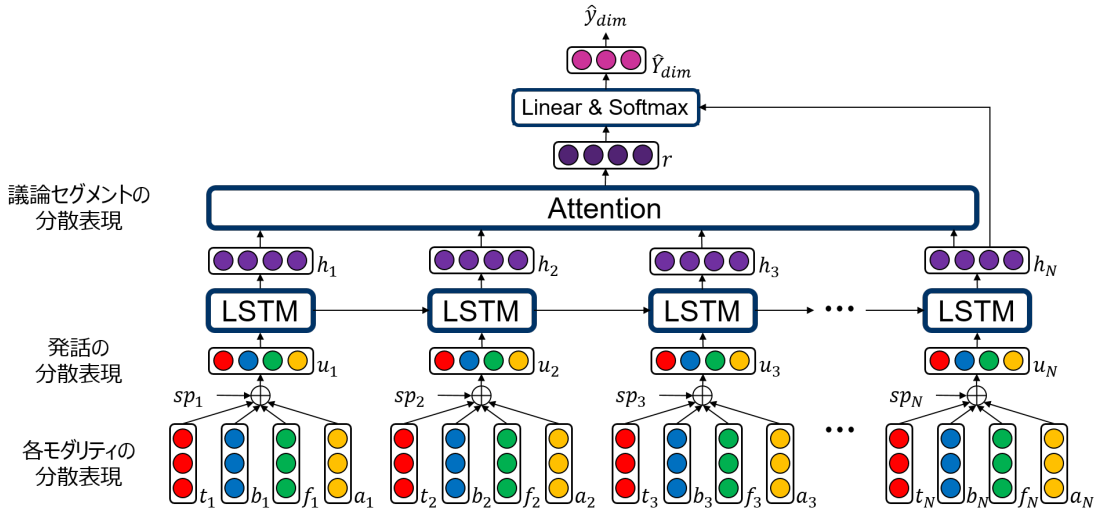


図 5.4 注意機構付き LSTM を用いた品質評価モデル 2

あった場合，前者と後者で話し手の自信の度合いが違うことを人間は感じることができる．一方，$\boldsymbol{t}_i$ のように単語の系列を工夫無しに 1 つのベクトルにしてしまった場合，このような単語の語順の情報を失ってしまう可能性がある．そこで，発話系列と単語系列という異なるレイヤーの系列情報を独立の LSTM によって取り扱う品質評価モデルを対話データを取り扱う研究 [46] を参考に構築する．

5.4.1 階層的 Long Short-Term Memory

図 5.5 に発話系列と単語系列を同時に考慮できる階層的 LSTM を用いた品質評価モデルを示す．図 5.2 と異なる点は，入力発話のテキスト情報を表現するベクトル $\boldsymbol{t}_i$ が，各単語のベクトルを LSTM によってエンコードしたベクトル $\boldsymbol{w}_i$ となっている点である．形式的には単語順を考慮した発話のテキスト情報を表現するベクトル $\boldsymbol{w}_i$ は以下のように計算される．

$$\boldsymbol{h}_{i,j}^{Uttr} = LSTM^{Uttr}(\boldsymbol{w}_{i,j}, \boldsymbol{h}_{i,j-1}^{Uttr}, \boldsymbol{c}_{i,j-1}^{Uttr}) \tag{5.17}$$

$$\boldsymbol{w}_i = \boldsymbol{h}_{i,M_i}^{Uttr} \tag{5.18}$$

$LSTM^{Uttr}()$ は単語ベクトルの系列をエンコードする発話 LSTM，$\boldsymbol{w}_{i,j}$ は議論セグメント S における i 番目の発話の j 番目の単語を表現するベクトル，$\boldsymbol{h}_{i,j}^{Uttr}$ は i 番目の発話の時刻 j における発話 LSTM の隠れ層，$\boldsymbol{c}_{i,j}^{Uttr}$ は i 番目の発話の時刻 j に

おける発話 LSTM のメモリセル，M_i は i 番目の発話の単語数をそれぞれ表す．そして，本モデルにおける発話ベクトル $\boldsymbol{u}_i$ は以下のような形で表される．

$$\boldsymbol{u}_i = [\boldsymbol{sp}_i; \boldsymbol{w}_i; \boldsymbol{b}_i; \boldsymbol{f}_i; \boldsymbol{a}_i] \tag{5.19}$$

これ以降，予測品質スコアの推定値算出までの処理は LSTM を用いた手法と同様に U を入力して，最終的な状態 $\boldsymbol{h}_N^{Cont}$ を獲得する．

$$\boldsymbol{h}_i^{Cont} = LSTM^{Cont}(\boldsymbol{u}_i, \boldsymbol{h}_{i-1}^{Cont}, \boldsymbol{c}_{i-1}^{Cont}) \tag{5.20}$$

$LSTM^{Cont}()$ は単語ベクトルの系列をエンコードする発話 LSTM，$\boldsymbol{h}_i^{Cont}$ は時刻 i における文脈 LSTM の隠れ層，$\boldsymbol{c}_i^{Cont}$ は時刻 i における文脈 LSTM のメモリセルをそれぞれ表す．そして，$\boldsymbol{h}_N$ を用いて出力ラベルの予測確率分布 $\hat{Y}_{dim}$ を求める．

$$\hat{Y}_{dim} = \mathrm{softmax}(\boldsymbol{W}_s \boldsymbol{h}_N^{Cont} + \boldsymbol{b}_s) \tag{5.21}$$

最後に，分布の最大値と対応するインデックスを獲得することで予測品質スコア $\hat{y}_{dim}$ を獲得する．

$$\hat{y}_{dim} = \arg\max_{y_{dim}} \hat{Y}_{dim} \tag{5.22}$$

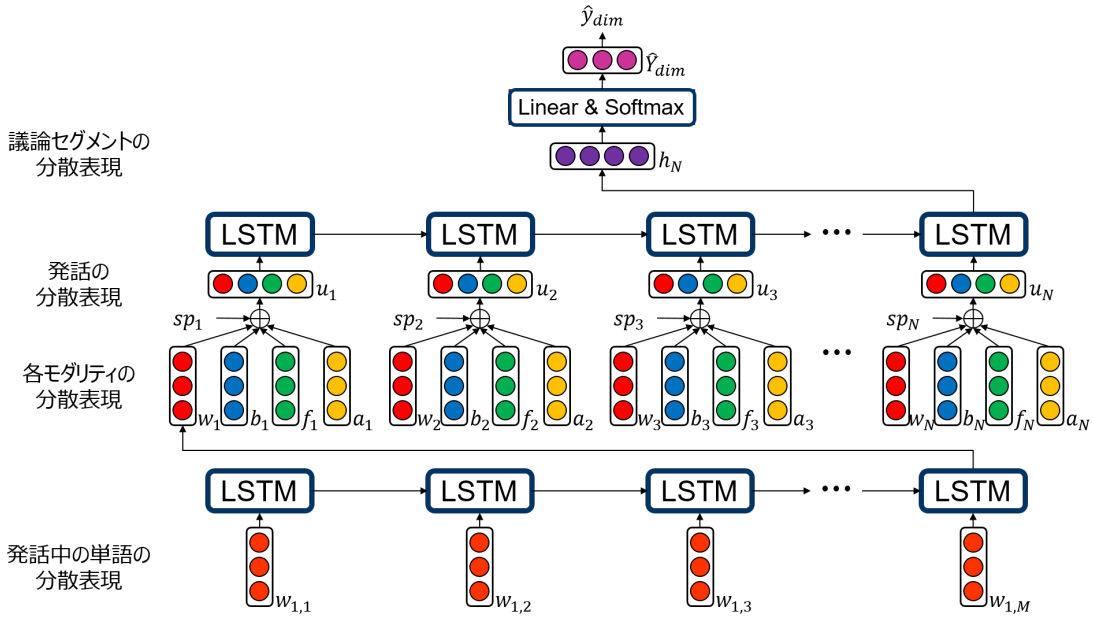


図 5.5 階層的 LSTM を用いた品質評価モデル

第6章　実験

本章では5章のモデルを3章から4章にかけて構築したデータセットに適用し，各モデルの自動評価の性能について報告する．各モダリティ情報の有無による性能について報告し，モダリティ情報の組み合わせによる評価値の変化や注意機構の可視化を基に結果の考察を行う．なお，本章以降では図5.1から図5.5のモデルをそれぞれ「SVM」「LSTM」「A-LSTM1」「A-LSTM2」「HLSTM」として参照する．

6.1　実験設定

本研究では，発話者の切り替わりを表す変数 $\boldsymbol{sp}_i$ は，時刻 i の発話の話者と時刻 $i-1$ の発話の話者が一致している場合は0とし，それ以外は1としている．モデルの入力となる発話のテキスト情報を表すベクトル $\boldsymbol{t}_i$ および $\boldsymbol{w}_{i,j}$ については，東北大学が公開しているBERT[47]の事前学習モデル[1]を用いて獲得する．本研究では先頭と末尾にそれぞれCLSトークンおよびSEPトークンを追加したものをBERTに入力し，[CLS]トークンに対応するBERTの11層目（768次元）を $\boldsymbol{t}_i$，各単語に対応するBERTの11層目（768次元）を $\boldsymbol{w}_{i,j}$ とした．発話の動作情報を表すベクトル $\boldsymbol{b}_i$ については，OpenPoseによって推定された発話中の上半身（鼻，首，右肩，右肘，右手首，左肩，左肘，左手首，右目，左目，右耳，左耳）のxy座標点をそれぞれフレームの進行方向に対して平均値および標準偏差で畳み込んだベクトル（48次元）とした．発話の表情情報を表すベクトル $\boldsymbol{f}_i$ については，OpenFaceによって推定された顔および目の特徴点のxy座標，視線方向，頭の位置と向き，および Facial Action Units (AUs) の有無の値をそれぞれフレームの進行方向に対して平均値および標準偏差で畳み込んだベクトル（586次元）とした．発話の音声情報を表すベクトル $\boldsymbol{a}_i$ については，Surfboardによって獲得した発話音声の13次元MFCC，RMS，基本周波数，スペクトル重心の値をそれぞれ時間の進行方向に対して最小値，最大値，平均値，および標準偏差で畳み込んだ値と，ジッタとシマの値を有するベクトル（72次元）とした．

[1] https://github.com/cl-tohoku/bert-japanese

表 6.1 実験で用いる正解スコアの分布

評価軸	略称	L	M	H
合理性	Re	13	89	76
有効性	Ef	9	97	72

4.2 節で構築した正解ラベルの信頼性をクリッペンドルフ α 係数で検証した．その中でも α 係数が相対的に高い，言い換えると 0.1 を超えている評価軸は「合理性（Re)」「充足性（GS)」「有効性（Ef)」「順序性（Ar)」の 4 つであった．これら 4 つの評価軸のスコア分布を表 4.5 で確認すると，主要評価軸の 2 つはデータの分布のバランスが良い一方，従属評価軸の 2 つは M のデータが多く偏りがある．そこで，本研究では主要評価軸の合理性（Re）と有効性（Ef）を用いて実験を実施する．更に，ラベルあたりのデータ数確保のために，主要評価軸のスコア VL および VH をそれぞれ L，H と見なすビニング処理を行い，データセットの再構築を行った．実験に用いるデータセットの統計値を表 6.1 に示す．

本研究における品質推定タスクはある議論セグメントに対して L，M，H の 3 つのタグを付与する 3 値分類問題として定式化できる．したがって，LSTM，A-LSTM1，A-LSTM2，HLSTM を学習する際の損失は L2 正則化付クロスエントロピー誤差を用い，この損失を最小にする方向で学習を実施した．

$$\mathcal{L} = -\sum_{s} \mathbb{1}(U, s) \log(\hat{Y}_{dim}) + \frac{\lambda}{2}\|\mathbf{W}\|^2 \tag{6.1}$$

s は品質スコア，$\mathbb{1}(U, s)$ は正解スコアに対応する要素が 1，それ以外の要素が 0 となるような正解確率分布，$\mathbf{W}$ はモデルの学習可能なパラメータ，λ は重み減衰に関するハイパーパラメータを表す．表 6.2 に本実験で利用した深層学習モデルのハイパーパラメータを示す．

最適化手法は確率的勾配降下法（SGD）[48] に慣性項（Momentum）[49] を付与した SGD Momentum を利用した．時刻 i におけるモデルの学習パラメータは以下のような式で決定される．

$$\mathbf{w}^{i+1} \leftarrow \mathbf{w}^{i} - \eta \frac{\partial E(\mathbf{w}^{i})}{\partial \mathbf{w}^{i}} + \alpha \Delta \mathbf{w}^{i} \tag{6.2}$$

η は学習係数，α は慣性項に関するハイパーパラメータを表す．損失計算，および最適化に必要なハイパーパラメータについては，表 6.2 にネットワークに関するパラメータとまとめて掲載している．なお，SVM の構築・実験には scikit-learn[50]

表 6.2 深層学習モデルのハイパーパラメータ

パラメータ	数値
LSTM の隠れ層（発話）	500
LSTM の隠れ層（文脈）	500
バッチサイズ	32
エポック数	50
学習率（η）	0.01
ドロップアウト	0.2
モーメンタム（α）	0.95
重み減衰（λ）	0.001

を，LSTM，A-LSTM1，A-LSTM2，HLSTM の構築・実験には PyTorch[51] を利用した．

本研究で構築したコーパスの収録対話数は 10 対話，議論セグメント数が 178 と少数である．そこで，訓練データ 8 対話，検証データ 1 対話，評価データ 1 対話という形で，データセット中の全ての対話が評価データとして用いられるよう組み合わせを作成し，10 対話交差検証を実施した結果を 1 実験の評価値とする．評価指標には分類問題の指標として用いられる F 値（F-measure）を利用した．また，結果の頑健性を保証するために実験を 5 回実施し，本論文ではそのマクロ平均値を報告し，議論する．

6.2 実験結果

まずはじめに，各議論セグメントの言語情報のみを利用した際の各モデルの品質評価性能について報告する．その後，言語情報に動作・表情・音声といった言語外の情報を利用した場合の品質評価性能についても報告する．これは，先行研究 [4] が指摘しているように，話し言葉の議論においては言語外の情報も品質に影響を与える可能性が存在するためである．

表 6.3 に言語情報のみを利用するユニモーダル設定における各モデルの品質評価性能を示す．baseline は全てのセグメントを全て M としたときの評価値である．Ave. は L，M，H における F 値のマイクロ平均である．太字で示されている評価値はそのカラムにおいて最大となった評価値，下線で示されている評価値はユニモーダル情報を利用した各モデルの間で最大となった評価値をである．実験の結

表 6.3 ユニモーダル設定における各モデルの評価性能

特徴量	モデル	Re				Ef			
		L	M	H	Ave.	L	M	H	Ave.
-	baseline	0.000	0.667	0.000	0.333	0.000	0.706	0.000	0.384
T	SVM	0.000	**0.611**	**0.340**	**0.451**	0.000	**0.635**	0.278	**0.459**
	LSTM	0.000	0.576	0.229	0.387	0.000	0.511	**0.369**	0.428
	A-LSTM1	0.000	0.589	0.276	0.412	0.000	0.590	0.275	0.433
	A-LSTM2	0.000	0.465	0.325	0.371	**0.040**	0.577	0.335	0.452
	HLSTM	0.000	0.520	0.230	0.359	0.000	0.580	0.352	**0.459**

果，Re ついては SVM の評価値が最大，Ef については SVM と HLSTM の評価値が最大となることを確認した．Re について詳しく観察すると，全ての品質ラベルに対する評価値において SVM の評価値が高い傾向にある．LSTM や A-LSTM1，A-LSTM2，HLSTM のような発話の系列を捉えるモデルの品質評価の性能がより高いことを予想していたが，Re の評価においては系列情報の有効性を確認することはできなかった．一方，評価軸 Ef について観察すると，深層学習によるモデルの評価値が相対的に SVM に近い，特に HLSTM においては等しい性能にありながら，品質ラベル H の値は SVM より高いことを確認した．つまり，評価軸 Ef の評価においては系列情報を用いることに一定の有効性があると結論付けられる．まとめると，議論セグメントの論理性など，議論内容そのものを評価する際には系列情報は有効ではない一方，議論内容のわかりやすさや情動性など，議論を聴いた人の感じ方に関する評価においては系列情報がある程度有効である可能性があることが示唆された．この結果を解釈すると，合理性に関する評価では議論全体で出現する発話の総合的な情報が必要である一方，有効性に関する評価ではセグメント内に出現する発話の順番が必要となることになる．

続いて，表 6.4 および表 6.5 に，表 6.3 のモデルに動作，顔，音声の特徴量を加えた各モデルの評価値を示す． Re については，モデルへと入力するモダリティの情報を拡張しても，SVM および深層学習のモデルにおいて品質推定性能の大きな変化は見られなかった．ユニモーダル，バイモーダル，マルチモーダルの設定における最高性能モデル（太字+下線）を比較すると，言語情報と顔情報を用いた HLSTM が最大性能（0.459）となった．しかし，この評価値は言語情報を用いた SVM の評価値（0.451）と 0.008 しか変わらず，それ以外のモデルについてはスコアが低くなっている．つまり，合理性（Re）に関する評価においては言語外による情報は有効ではない可能性があると考えられる．一方，Ef についての結果を確

表 6.4 バイモーダル設定における各モデルの評価性能

特徴量	モデル	Re				Ef			
		L	M	H	Ave.	L	M	H	Ave.
-	baseline	0.000	0.667	0.000	0.333	0.000	0.706	0.000	0.384
TB	SVM	**0.040**	0.549	0.143	0.338	0.000	**0.679**	0.029	0.382
	LSTM	0.000	0.587	**0.245**	**0.398**	0.000	0.598	**0.376**	**0.478**
	A-LSTM1	0.000	**0.589**	0.228	0.392	0.000	0.596	0.359	0.470
	A-LSTM2	0.000	0.538	0.161	0.337	0.000	0.553	0.274	0.412
	HLSTM	0.000	0.524	0.216	0.354	0.000	0.576	0.294	0.433
TF	SVM	0.000	0.541	0.156	0.337	0.000	**0.671**	0.042	0.383
	LSTM	0.000	**0.613**	0.199	0.392	0.000	0.573	**0.311**	**0.438**
	A-LSTM1	0.000	0.545	0.269	0.387	0.000	0.572	0.282	0.426
	A-LSTM2	0.000	0.545	0.300	0.400	0.000	0.568	0.263	0.416
	HLSTM	0.000	0.535	**0.446**	**0.459**	0.000	0.561	0.273	0.416
TA	SVM	**0.115**	0.502	0.196	0.343	**0.127**	**0.618**	0.230	0.436
	LSTM	0.000	0.544	0.253	0.380	0.000	0.606	**0.355**	**0.476**
	A-LSTM1	0.000	**0.602**	0.230	0.399	0.000	0.598	0.352	0.468
	A-LSTM2	0.000	0.569	**0.318**	**0.420**	0.000	0.579	0.298	0.436
	HLSTM	0.000	0.589	0.218	0.388	0.000	0.500	0.263	0.379

認すると，SVM 以外については入力モダリティの種類が増加するごとに品質推定の性能が増加する傾向がいくつか見られることを確認した．特に，全てのモダリティを用いた A-LSTM1 が最も高い評価値（0.490）となった．つまり，Ef に関する評価においては，言語情報のみならず，その発言者の動作や表情，声のトーンなど様々な情報を用いることが重要であると考えられる．SVM の性能が例外的に低下した原因については，拡張されたモダリティの系列情報を平均と標準偏差のみでは表現できなかったためであると推測する．この実験結果については，Re と Ef それぞれの定義と比較すると一貫していることがわかる．つまり，Re（合理性）は議論の容認性，関連性，充足性といった議論の内容そのものに関連する評価軸であり，言語的な内容以外の要因では基本的に評定値が前後しないものである．そのため，各モデルに言語以外のモダリティ情報を導入しても合理性に関する品質評価の性能は上がらない，もしくは入力された情報がノイズとなり性能の低下を招くと考えられる．一方，Ef（有効性）は議論の信用性や情動性といった聴衆側の感情や，明瞭性，妥当性，順序性といった議論の理解しやすさを表す評定軸である．一般に，人に信用してもらえるように物事を伝えたり，人に共感などの感情

表 6.5 マルチモーダル設定における各モデルの評価性能

特徴量	モデル	Re				Ef			
		L	M	H	Ave.	L	M	H	Ave.
-	baseline	0.000	0.667	0.000	0.333	0.000	0.706	0.000	0.384
TBF	SVM	0.000	0.524	0.167	0.333	0.000	**0.648**	0.075	0.384
	LSTM	0.000	0.523	**0.349**	0.410	0.000	0.605	0.339	**0.467**
	A-LSTM1	0.000	0.514	0.268	0.371	0.000	0.565	**0.351**	0.450
	A-LSTM2	0.000	0.554	0.296	0.404	0.000	0.594	0.295	0.443
	HLSTM	0.000	**0.569**	0.306	**0.415**	0.000	0.508	0.339	0.414
TBA	SVM	0.000	0.500	0.156	0.317	0.000	**0.684**	0.049	0.392
	LSTM	0.000	0.430	**0.384**	0.379	0.000	0.563	**0.408**	**0.472**
	A-LSTM1	0.000	0.532	0.218	0.359	0.000	0.546	0.244	0.396
	A-LSTM2	0.000	0.528	0.258	0.374	0.000	0.518	0.352	0.425
	HLSTM	0.000	**0.576**	0.241	**0.391**	0.000	0.612	0.262	0.440
TFA	SVM	0.000	0.532	0.127	0.320	0.000	**0.684**	0.082	0.406
	LSTM	0.000	0.565	0.248	0.388	0.000	0.623	**0.362**	**0.486**
	A-LSTM1	0.000	0.605	0.224	**0.398**	0.000	0.606	0.281	0.444
	A-LSTM2	0.000	0.510	**0.310**	0.387	0.000	0.552	0.247	0.401
	HLSTM	0.000	**0.613**	0.148	0.370	0.000	0.549	0.327	0.431
TBFA	SVM	**0.116**	0.492	0.200	0.340	0.000	**0.693**	0.005	0.379
	LSTM	0.000	0.558	0.190	0.360	0.000	0.543	0.343	0.435
	A-LSTM1	0.000	**0.579**	0.255	0.398	0.000	0.627	0.365	**0.490**
	A-LSTM2	0.000	0.548	0.152	0.339	0.000	0.508	0.315	0.404
	HLSTM	0.000	0.552	**0.303**	**0.405**	0.000	0.541	**0.386**	0.451

を求める場合には目を合わせるなど言外で何か工夫を行うことが考えられる．また，人は何か物事をわかりやすく説明する場合にはボディランゲージを利用することも考えられる．そのため，有効性の品質評価の性能は言語・非言語に関する情報を同時に導入することで向上すると結論づけられる．

次に，注意機構付きの深層学習モデルが議論セグメントのどの発話に着目して品質推定を行っているかを注意機構の重みという側面から分析する．図 6.1 に A-LSTM1 モデルに言語情報（T）のみを入力して合理性（Re）の品質推定を行った際に出力された重みの例を示す．このモデルを選んだ理由については合理性を評価する注意機構付きの手法の中で最も評価値の高かったからである．このセグメントは対話 20191120_T1（成人の拳銃所持・携帯の権利を認めるべきである）に属す

るセグメントの 1 つで，合理性のラベルが H とされているセグメントである．図 6.1 の議論では拳銃の所持を認めないグループの話者が世の中には人の命を奪いかねない道具は沢山存在するが，それらの道具は利便性が存在するから利用しているのに対し，拳銃は利便性がないことを主張している．このとき，その主張の内容に関連する発話にアテンションの重みが付与されることを期待していたが，本研究で確認した結果では「うん」という相槌の発話に対して重みが割り当てられていることがわかる．実際にモデルが正確に品質を推定できたセグメントについても確認すると，セグメントの要点がまとめられている発話よりも,「うん」といった相槌や比較的文字列長の短い発話に重みが割り当てられていることを確認した．つまり，言語情報を用いた A-LSTM のモデルはセグメント内で話を聴いている側の反応，つまり議論の空気感を基にラベルを付与した可能性が存在する．一方でアテンションの解釈可能性，つまりモデルの推論過程が人に理解できる形で表現されているかという点については様々な議論が行われており [52][53][54]，必ずしもこの結果が上記の解釈と一致しているとは限らない．また，有効性を評価する注意機構付きの手法で最も評価値の高かった A-LSTM1 についても解釈を試みたが，規則性などを発見することはできなかった．

話者	発話	
C	他の殺傷能力が高いやつ	
C	って	
D	うん	
C	なんかまあ	
C	ナイフだとか	
D	うんうんうん	
C	あとはまあ	
C	車だとか	
D	はいはい	
B	車<笑>	
C	あ車まあまあまあはい	
A	まあまあまあまあまあ	
D	まあまあまあまあまあ	
C	後ダイナマイトとか	
D	うん	
C	まあ色々あると思うんですけど	
D	うん	
C	今挙げた車ナイフ	
C	ダイナマイトって	
C	殺傷能力よりも	
C	人間の生活の利便性の方が勝ってるから	
C	なんか使用が許可されてる	
C	と思うんですよ	
D	うんうんうん	

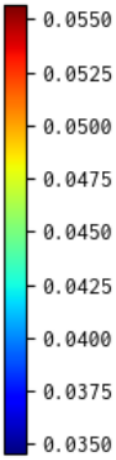


図 6.1 A-LSTM1（T）のアテンションの可視化

第7章　おわりに

本論文では複数人議論の自動品質評価を目的として，複数人議論コーパスの構築を実施し，そのデータを対象としたと議論の自動評価モデルを構築，性能評価を実施した．1章および2章では本研究の背景や関連研究の紹介と，それらと比較した際の本研究の立ち位置について述べた．

3章では討論と合意形成という2種類の対話シナリオと5つの命題を設定し，合計10対話（約200分）の対話収録を実施した．複数のカメラおよびピンマイクで獲得した映像・音声データを同期し，対話メディアデータを獲得した．その後，個人を特定できる情報を削除し，かつ機械上でデータの取り扱いを容易にするためのデータ構造化を実施した．収録メディアデータを基に，アノテーションツールELANを用いて，0.2秒の無声区間を区切りとする転記単位で発話の書き起こしを実施した．本研究において構築したコーパスの発話数は7,449発話となった．また，各議論参加者の骨格，手，顔，視線，頭部の情報をOpenPose，およびOpenFaceを用いて獲得した．さらに，各発話に対応する音声ファイルから声道特性や声質に関する特徴量をSurfboardを用いて獲得した．その後，書き起こした発話を話の意味的なまとまりで分割するトピックセグメンテーションを実施した．

4章では2種類の品質評定手法を設計し，それぞれの評定結果について報告した．1つめはエッセイの自動評価に関する研究を参考に，各議論参加者の陳述内容の評定を協力者に依頼した．この実験を通して，陳述が意味に基づいて分割されない点や，他の議論参加者との相互作用が考慮できていない点が問題となることを明らかにした．そこで，2つめのアプローチとしてセグメンテーションされた対話データの品質を議論の品質推定の理論を参考に評定する作業を依頼した．こちらの手法においても十分な一致率のラベルは獲得できなかったが，一定の品質が保証されていることを定量的に示した．

5章では議論セグメントの発話系列を入力として受け取り，議論セグメントの品質を推定するモデルを提案した．はじめに，少量のデータでも学習が可能なSVMを用いた手法を提案した．SVMでは発話の系列を統計量によって畳み込むため十分な文脈情報を考慮できない可能性を踏まえ，発話の系列情報を考慮できる回帰型ニューラルネットワークを用いた手法，および注意機構を用いて議論セグメントのより重要な情報を重点的に処理する注意機構付きの回帰型ニューラルネット

ワークを用いた手法を提案した．更に，発話の単語系列情報についても考慮する階層的回帰型ニューラルネットワークによる手法も提案した．

6 章では 5 章のモデルの評価性能を 3 章および 4 章で構築したデータセットで検証した．実験の結果，現在のデータセットを用いた提案手法では，議論の合理性は F 値で 0.459，有効性は 0.490 の性能で自動評価できることを確認した．また，議論の合理性は議論内容に関連する評価軸であるため言語以外のモダリティ情報を用いても評価性能が向上しない一方，有効性では動作や音声といった複数のモダリティ情報を同時に入力することで性能向上が見込めることを実験によって確認した．

最後に，本研究における今後の課題，および展望について以下にまとめる．

1. コーパスの収録対話数の拡張
   本研究では新型コロナウイルスの感染拡大を受け，十分な対話数を収録したコーパスの構築を実現することができなかった．そのため今後は，現在収録している対話数と同等量以上の対話収録を実施することが喫緊の課題としてあげられる．その際に議論参加者のサーベイの時間を省略するなどの工夫を施し，様々な品質の議論を収録することも必要になってくると著者は考えている．

2. 議論セグメントの品質評定手法の改善
   また，現在の議論の品質推定に関する研究では，質の高い正解ラベルを構築する手法が確立されていない．そのため，評価の際にチェックリスト方式を導入することや，評価軸の定義を簡易化するなどしてより信頼性の高い評価ラベルを付与する手続きの構築が必要である．

3. 品質評価手法の高度化と拡張
   本研究では合理性および有効性に関するスコアを提案モデルによって推定する実験を実施し，合理性と有効性それぞれ F 値で 0.459，0.490 程度の精度で品質推定が可能であることを示したが，これらのスコアには改善の余地がある．議論の中の発話文脈と各発話者の状態を独立に考慮し，対話の状態や各発話の状態をより高品質に生成するモデル [55][56] が提案されているため，これらのモデルの利用による品質推定性能の確認が 1 つのアプローチとして考えられる．また近年では，訓練データ以外の外部知識を用いてモデルの性能を向上されている研究が行われており，本研究が取り組むタスクにおいても議論に関する知識グラフ [57] を用いて評価モデルの改良などのアプローチも考えられる．さらには，本研究では取り扱うことができなかった議論セグメント内に存在する論について適切性の軸から評価する課題についても取り組む必要がある．

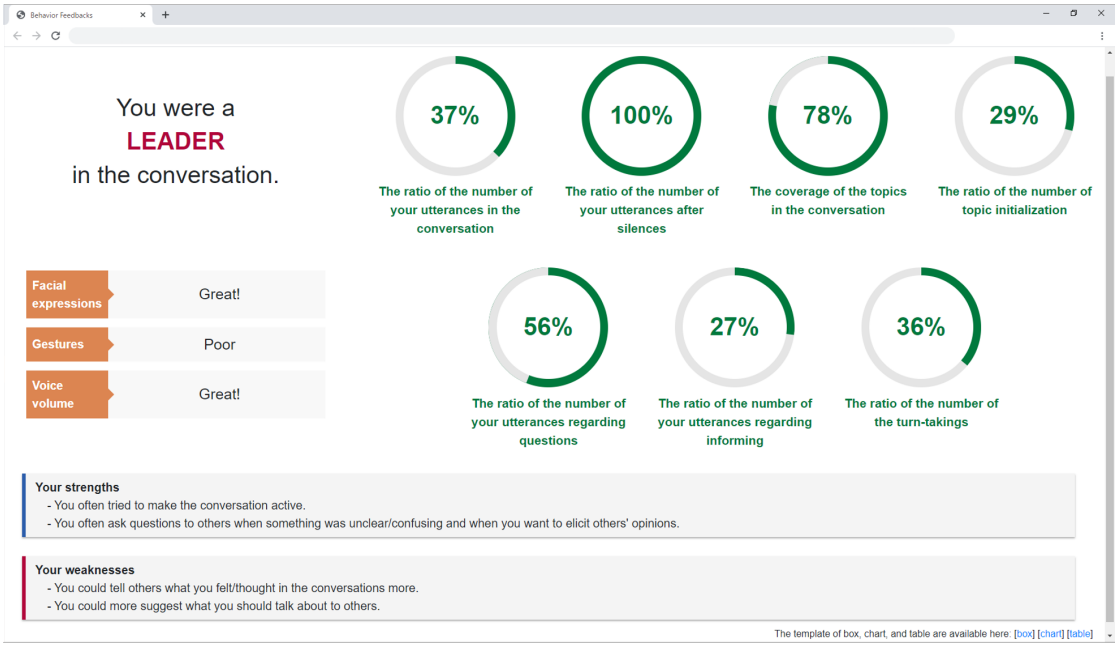


図 7.1 振る舞いフィードバックシステムのプロトタイプ

4. 評価レポートシステムの作成
本研究の最終目的は議論状態の可視化による議論参加者の学習や評価者の評価活動の支援である．著者はこれまでに，議論においてリーダー的役割を果たす人物とそうでない人物を識別するモデルの構築と，その人物の行動パターンを分析した情報をフィードバックするプロトタイプシステムを開発している [58]（図 7.1）．本研究において構築した品質評価モデルによる結果なども含めて，情報を総合的かつリアルタイムに近い速度でフィードバックを行えるシステムを開発することも課題の 1 つである．

謝辞

本研究を遂行し学位論文を作成するにあたり，様々な面で御指導，御鞭撻を賜りました，指導教官である嶋田和孝教授に深く感謝いたします．在籍していた 3 年（留学期間も含めると 4 年）の間,「もっと大きくて有名な研究室でハイインパクトな研究を...」など数々の失言の繰り返す著者に対して沢山の経験や学びの機会を与えてくださったこと大変嬉しく思っております．研究室在籍中に多くの学会/研究会における発表や西日本新聞社との共同研究など様々な活動を通じて，改めて研究活動やものづくりの楽しさを知ることができました．また，コロナ禍であるにも関わらず第一志望の研究所に内定を頂くこともできました．そして，本論文執筆中に論文誌の採録と NLC 研究会の学生研究賞受賞という大きな実績を残すことができるまで成長させていただきました．これも全て生意気な著者に対して懐深く熱心な指導をしてくださいました先生のおかげだと思っております．本当にありがとうございます．

加えて，中間報告会などで多くのご助言を下さいました，乃万司教授，岡部孝弘教授，山本邦雄助教，ならびに計算機など研究環境に関してご助言を下さいました技術職員の松元隆二さんに感謝いたします．特に，岡部孝弘教授には御厚意で岡部研究室で開催される深層学習勉強会への参加を許可して頂き，専門知識について深く学べる環境を設けていただいたことを改めてお礼申し上げます．

また，本研究を遂行するにあたり，ディスカッションやアノテーションなど様々な研究活動にご協力を下さいました嶋田研究室の皆様に感謝いたします．特に，著者が嶋田研究室に配属されて以降，論文執筆から技術的な支援まで在学期間中嫌な顔一つせず沢山ご協力いただきました卒業生の山村崇先輩に深く感謝いたします．また，本研究の遂行に必要となる技術の指導や執筆作業にご自身の学位取得でお忙しい中ご協力いただきました博士課程の肥合智史先輩に感謝いたします．加えて，著者の卒業研究のテーマを継続し，有意義な結果を共に導き出してくれた卒業生の本多幸希君にもこの場を借りてお礼申し上げたいと思います．

最後に，25 年間著者の勉学・研究活動を精神的にも経済的にも支えてくれた家族に感謝します．ありがとうございました．すねをかじりすぎたので来年度からは経済的に自立し，親孝行を頑張りたいと思います．

参考文献

[1] James J. Heckman and Tim Kautz. Hard Evidence on Soft Skills. *Labour Economics*, Vol. 19, No. 4, pp. 451 – 464, 2012.

[2] Zixuan Ke and Vincent Ng. Automated Essay Scoring: A Survey of the State of the Art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 6300–6308, 2019.

[3] Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, 2003.

[4] 武川直樹, 中山知大, 徳永弘子, 大和淳司, 山下直美. グループディスカッションにおける発言者の言語/非言語の表出と評価者評価の関係の分析. 電子情報通信学会論文誌 D, Vol. 101, No. 2, pp. 284–293, 2018.

[5] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The AMI meeting corpus: A pre-announcement. In *Proceedings of the 2nd International Workshop on Machine Learning for Multimodal Interaction (MLMI)*, pp. 28–39, 2005.

[6] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. The ICSI Meeting Corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 364–367, 2003.

[7] Nick Campbell, Toshiyuki Sadanobu, Masataka Imura, Naoto Iwahashi, Suzuki Noriko, and Damien Douxchamps. Multimedia Database of Meetings and Informal Interactions for Tracking Participant Involvement and Discourse Flow. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pp. 391–394, 2006.

[8] Oya Aran, Hayley Hung, and Daniel Gatica-Perez. A Multimodal Corpus for Studying Dominance in Small Group Conversations. In *Proceedings of the 7th International Conference on Language Resources and Evaluation Workshop on Multimodal Corpora (LREC)*, pp. 2223–2232, 2010.

[9] Maria Koutsombogera and Carl Vogel. Modeling Collaborative Multimodal Behavior in Group Dialogues: The MULTISIMO Corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pp. 2945–2951, 2018.

[10] Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. Conversational Flow in Oxford-style Debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 136–141, 2016.

[11] Volha Petukhova, Andrei Malchanau, Youssef Oualil, Dietrich Klakow, Saturnino Luz, Fasih Haider, Nick Campbell, Dimitris Koryzis, Dimitris Spiliotopoulos, Pierre Albert, Nicklas Linz, and Jan Alexandersson. The Metalogue Debate Trainee Corpus: Data Collection and Annotations. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pp. 749–755, 2018.

[12] 林佑樹, 二瓶芙巳雄, 中野有紀子, 黄宏軒, 岡田将吾. グループディスカッションコーパスの構築および性格特性との関連性の分析. 情報処理学会論文誌, Vol. 56, No. 4, pp. 1217–1227, 2015.

[13] Christopher Olshefski, Luca Lugini, Ravneet Singh, Diane Litman, and Amanda Godley. The Discussion Tracker Corpus of Collaborative Argumentation. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pp. 1033–1043, 2020.

[14] France H. van Eemeren and A. Francisca Snoeck Henkemans. 議論学への招待 建設的なコミュニケーションのために. 大修館書店, 2018.

[15] Isaac Persing and Vincent Ng. Modeling Argument Strength in Student Essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and the 7th International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 543–552, 2015.

[16] Zahra Rahimi, Diane Litman, Elaine Wang, and Richard Correnti. Incorporating Coherence of Topics as a Criterion in Response-to-Text Assessment of the Organization of Writing. In *Proceedings of the 10th workshop on Innovative Use of NLP for Building Educational Applications*, 2015.

[17] Elena Cabrio and Serena Vilata. Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012.

[18] Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. Argumentation Quality Assessment: Theory vs. Practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 250–255, 2017.

[19] Ivan Habernal and Iryna Gurevych. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1589–1599, 2016.

[20] Ivan Habernal and Iryna Gurevych. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1214–1223, 2016.

[21] Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. Efficient Pairwise Annotation of Argument Quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5772–5781, 2020.

[22] Henning Wachsmuth, Benno Stein, and Yamen Ajjour. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 1117–1127, 2017.

[23] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, 1999.

[24] Shogo Okada, Yoshihiko Ohtake, Yukiko Nakano, Hayashi Yuki, Hung-Hsung Huang, Yutaka Takase, and Katsumi Nitta. Estimating Communication Skills Using Dialogue Acts and Nonverbal Features in Multiple Discussion Datasets. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*, pp. 169–176, 2016.

[25] Wen Dong and Alex "Sandy" Pentland. Quantifying group problem solving with stochastic analysis. In *In Proceedings of International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI)*, pp. 1–4, 2010.

[26] Umut Avci and Oya Aran. Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction. *IEEE Transactions on Multimedia*, Vol. 18, No. 4, pp. 643–658, 2016.

[27] Gabriel Murray and Catharine Oertel. Predicting Group Performance in Task-Based Interaction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, pp. 14–20, 2018.

[28] Go Miura and Shogo Okada. Task-independent multimodal prediction of group performance based on product dimensions. In *Proceedings of the 21st ACM International Conference on Multimodal Interaction (ICMI)*, p. 264?273, 2019.

[29] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 176–187, 2017.

[30] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. ELAN: A Professional Framework for Multimodality Research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pp. 1556–1559, 2006.

[31] 小磯花絵, 土屋菜穂子, 間淵洋子, 斉藤美紀, 籠宮隆之, 菊池英明, 前川喜久雄. 「日本語話し言葉コーパス」における書き起こしの方法とその基準について. 日本語科学, Vol. 9, pp. 43–58, 2001.

[32] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, Vol. 43, No. 1, pp. 172–186, 2021.

[33] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *Proceedings of 13th IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pp. 59–66, 2018.

[34] Raphael Lenain, Jack Weston, Abhishek Shivkumar, and Emil Fristed. Surfboard: Audio Feature Extraction for Modern Machine Learning. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.

[35] Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse Segmentation of Multi-Party Conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 562–569, 2003.

[36] Takashi Yamamura, Kazutaka Shimada, and Shintaro Kawahara. The Kyutech Corpus and Topic Segmentation Using a Combined Method. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR)*, pp. 95–104, 2016.

[37] 冨山健, 二瓶芙巳雄, 高瀬裕, 中野有紀子. マルチモーダル特徴量を用いた談話セグメントの検出. 人工知能学会全国大会論文集, No. 4F3OS11b02, 2019.

[38] Weiqun Xu, Jean Carletta, Jonathan Kilgour, and Vasilis Karaiskos. Coding Instructions for Topic Segmentation of the AMI Meeting Corpus Version 1.1, 2005.

[39] Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 621–631, 2018.

[40] Aristotle. *On Rhetoric: A Theory of Civic Discourse*. Oxford University Press, 2007.

[41] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer science & business media, 2013.

[42] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[43] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[44] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 606–615, 2016.

[45] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 207–212, 2016.

[46] Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. A Hierarchical Neural Model for Learning Sequences of Dialogue Acts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 428–437, 2017.

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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 (NAACL-HLT)*, pp. 4171–4186, 2019.

[48] Léon Bottou. Stochastic Gradient Learning in Neural Networks. In *Proceedings of Neuro-Nîmes 91*, 1991.

[49] Ning Qian. On the Momentum Term in Gradient Descent Learning Algorithms. *Neural Networks*, Vol. 12, No. 1, pp. 145 – 151, 1999.

[50] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron

Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research (JMLR)*, Vol. 12, pp. 2825–2830, 2011.

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pp. 8024–8035, 2019.

[52] Sofia Serrano and Noah A. Smith. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2931–2951, 2019.

[53] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 3543–3556, 2019.

[54] Sarah Wiegreffe and Yuval Pinter. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

[55] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 6818–6825, 2019.

[56] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 154–164, 2019.

[57] Khalid Al-Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. End-to-End Argumentation Knowledge Graph Construction. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 7367–7374, 2020.

[58] Tsukasa Shiota, Kouki Honda, Kazutaka Shimada, and Takeshi Saitoh. Development and Application of Leader Identification Model Using Multimodal Information in Multi-party Conversations. *International Journal of Asian Language Processing (IJALP)*. (in press).

付録

A. 陳述評価におけるスコア基準一覧

表 7.1 説得力採点基準

スコア	概要
6	とても力強く明確な陳述．彼/彼女の陳述を聴いた人のほとんどが納得すると考えられる．
5	力強く明確な陳述．彼/彼女の陳述を聴いた人の過半数が納得すると考えられる．
4	理解できる陳述．彼/彼女の陳述を聴いた人の半数程度が納得すると考えられる．
3	概ね理解できる陳述．彼/彼女の陳述を聴いた人が元々同じ考えなら納得すると考えられる．
2	理解し難い陳述．彼/彼女の陳述を聴いた人の過半数が納得しないと考えられる．
1	とても理解し難い陳述．彼/彼女の陳述を聴いた人のほとんどが納得しないと考えられる．

表 7.2 発想力採点基準

スコア	概要
6	とても独創的かつ具体的である．陳述をより向上するための内容の改善がほとんどない．
5	独創的かつ具体的である．内容の不明瞭さが僅かにあるが独創的な発想で具体性が十分ある．
4	概ね独創的かつ具体的である．独創性が秀でて高くないが，陳述の内容は練られえている．
3	あまり独創的かつ具体的でない．独創性が高くなく，陳述内容に対して不明瞭な点が残る．
2	独創的かつ具体的でない．陳述内容の独創性が低く，曖昧で不明瞭である．
1	とても独創的かつ具体的でない．陳述内容が抽象的で曖昧性が高く独創性もない．

表 7.3 雄弁さ採点基準

スコア	概要
6	明確かつ言葉巧みに主張を行っており，不明瞭な陳述がない．陳述の理解を低下させるような言い直し，言い淀みがなく，強引な畳み掛けや詭弁のような言論がない．
5	明確に主張を行っているが，稀に不明瞭な陳述がある場合もある．陳述の理解を低下させるような言い直し，言い淀みがなく，強引な畳み掛けや詭弁のような言論がない．
4	上手く主張を行っているが，稀に不明瞭な陳述がある場合もある．陳述の理解を低下させるような言い直し，言い淀み，または弱い畳み掛けや詭弁のような言論がたまにある．
3	概ね上手く主張を行っているが，不明瞭な陳述がしばしばある．陳述の理解を低下させるような言い直し，言い淀み，または弱い畳み掛けや詭弁のような言論がたまにある．
2	あまり上手く主張を行えておらず，不明瞭な陳述がしばしばある．陳述の理解を低下させるような言い直し，言い淀み，または強引な畳み掛けや詭弁のような言論がある．
1	うまく主張を行えておらず，陳述が多くの場合不明瞭である．陳述の理解を低下させるような言い直し，言い淀み，または強引な畳み掛けや詭弁のような言論がある．

表 7.4 一貫性採点基準

スコア	概要
6	とても一貫性がある．陳述の構成がとても明確で，2 転 3 転することなく首尾貫徹している．
5	一貫性がある．陳述の構成が明確で，陳述が 2 転 3 転することがほぼ無く一貫している．
4	概ね一貫性がある．陳述の構成が概ね明確で，陳述があまり 2 転 3 転することがない．
3	あまり一貫性がない．陳述の構成が稀に崩れており，陳述が 2 転 3 転することがある．
2	一貫性がない．陳述の構成がしばしば崩れており，陳述が 2 転 3 転することがよくある．
1	とても一貫性がない．陳述の構成が多くの場合崩れており，陳述が一貫していない．

表 7.5 批判的思考力採点基準

スコア	概要
6	とても批判的思考力がある．主張や質疑応答等全ての文脈において客観性と合理性の高い陳述を行えている．
5	批判的思考力がある．主張や質疑応答等全ての文脈においてある程度客観性と合理性の高い陳述を行えている．
4	概ね批判的思考力がある．多くの文脈において客観性と合理性の高い陳述を行えている．
3	あまり批判的思考力がない．一部の文脈においてのみ客観性と合理性の高い陳述を行えている．
2	批判的思考力がない．ほとんどの文脈において客観性と合理性の高い陳述を行えていない．
1	とても批判的思考力がない．全ての文脈において客観性と合理性の高い陳述を行えていない．

B. 陳述評価におけるスコア基準一覧

合理性 (Re)

「合理性の高い議論」とは，議論を聴いている人が十分に容認できる方法で議題の解決に貢献している議論のことを指します．この overall な評価軸は主に議論の内容やその進行方法に関する評価軸です．以下の GA から GS を評価し，最終的な overall スコアを決定します．

容認性 (GA) 「議論の容認性が高い」とは，議論を聴いている人がその議論の内容と進行方法の両方を許容できる状態のことを指します．つまり，ここでは評価対象となっている議論の内容が対話の議題から逸脱しておらず適切か，そしてその議論の進行（各自の話の進め方/四人の話し合い方）が適切で許容できるかを評価してください．

関連性 (GR) 「議論の関連性が高い」とは，議論の内容が対話の議題の解決に向けて貢献している状態を指します．つまり，ここでは評価対象となっている議論における内容が対話の議題と十分関係した内容であるかを評価してください．

充足性 (GS) 「議論の充足性が高い」とは，議論の内容が，それに対する反論/批判を予測できている（客観的で批判的である）状態のことを指します．つまり，ここでは評価対象となっている議論の内容がそれに対して考えられる反論/批判などをある程度考慮し，客観性の高いものであるか（反対意見などに言及している，あるいは議論を聴いた人が簡単に反論/批判できないように練られているか）を評価してください．

有効性 (Ef)

「有効性の高い議論」とは，対話の議題に対する意見を聴き手（第三者）に納得/同意させる議論のことを指します．この overall な評価軸は主に議論の内容だけでなく，それに付随する修辞的な効果や物の伝え方の上手さ等に関する評価軸です．以下の Cr から Ar を評価し，最終的な overall スコアを決定します．

信用性 (Cr) 「議論の信頼性が高い」とは，議論を聴いている人が議論参加者の述べている内容/議論参加者自体を信頼するに値すると思える状態を指します．つまり，ここでは評価対象となっている議論で述べられている内容がどの程度信じられるかを評価してください．

情動性 (Em) 「議論の情動性が高い」とは，議論を聴いている人が議論参加者の述べている内容に対してよりオープンになる感情を作っている状態を指します．つまり，ここでは評価対象となっている議論で述べられている内容にオープンマインドになれる（一理あるな，と思い議論の内容を積極的に受け入れようと思える）かを評価してください．

明瞭性 (Cl) 「議論の明瞭さが高い」とは，議論において曖昧/抽象的な単語や言い回し，必要以上に難しい説明，議題からの逸脱が含まれていない状態を指します．つまり，ここでは評価対象となっている議論で述べられている内容がどの程度理解しやすいかを評価してください．

妥当性（Ap） 「議論の妥当性が高い」とは，議論における内容が逸脱していないだけでなく，議論参加者の用いる言語/態度がその議論に対する信頼性，議論を理解しようとする気持ちの創造を支持している状態を指します．つまり，ここでは評価対象となっている議論における言葉遣いや物の言い方が攻撃的/感情的でなく建設的であるかを評価してください．

順序性（Ar） 「議論の順序性が高い」とは，議論における論点，根拠，結論を正しい順番で示すことのできている状態を指します．つまり，ここでは評価対象となっている議論における論点と根拠，それから導出される結論までの流れが理解しやすい形で示されているかを評価してください．

主要評価軸評価値決定ルール

【合理性 (Re)】
if 従属評価軸の評価値のうち，2 つ以上が L のとき：
　if もし残りの 1 つが L のとき：
　　→ Very low（L,L,L）
　elif もし残りの 1 つが M のとき：
　　→ Very low または Low（L,L,M）
　else：
　　→ Low（L,L,H）
elif 従属評価軸の評価値のうち，2 つ以上が M のとき：
　if もし残りの 1 つが L のとき：
　　→ Low または Middle（M,M,L）
　elif もし残りの 1 つが M のとき：
　　→ Middle（M,M,M）
　else：
　　→ Middle または High（M,M,H）
elif 従属評価軸の評価値のうち，2 つ以上が H のとき：
　if もし残りの 1 つが L のとき：
　　→ High（H,H,L）
　elif もし残りの 1 つが M のとき：
　　→ High または Very high（H,H,M）
　else：
　　→ Very high（H,H,H）
else（従属評価軸の評価値がバラバラのとき）:
　→ Middle（L,M,H）

【有効性 (Ef)】
if 従属評価軸の評価値のうち，4 つ以上が L のとき：
→ Very low（L,L,L,L,*）
elif 従属評価軸の評価値のうち，4 つ以上が M のとき：
→ Middle（M,M,M,M,*）
elif 従属評価軸の評価値のうち，4 つ以上が H のとき：
→ Very high（H,H,H,H,*）
elif 従属評価軸の評価値のうち，3 つ以上が L のとき：
if 残りの 2 つが M,M のとき：
→ Very low または Low（L,L,L,M,M）
elif 残りの 2 つが M,H または H,H のとき：
→ Low（L,L,L,M(H),H）
elif 従属評価軸の評価値のうち，3 つ以上が M のとき：
if 残りの 2 つが L,L のとき：
→ Low または Middle（M,M,M,L,L）
elif 残りの 2 つが H,H のとき：
→ Middle または High（M,M,M,H,H）
else（それ以外）:
→ Middle（M,M,M,*,*）
elif 従属評価軸の評価値のうち，3 つ以上が H のとき：
if 残りの 2 つが L,L または L,M のとき：
→ High（L,L(M),H,H,H）
elif 残りの 2 つが M,M のとき：
→ High または Very high（M,M,H,H,H）
else（従属評価軸の評価値のうち，3 つ以上同じスコアがないとき）:
if 従属評価軸の評価値のうち，L と M が 2 つずつのとき：
→ Low または Middle（L,L,M,M,H）
elif 従属評価軸の評価値のうち，L と H が 2 つずつのとき：
→ Middle（L,L,M,H,H）
else（従属評価軸の評価値のうち，M と H が 2 つずつのとき）:
→ Middle または High（L,M,M,H,H）

業績リスト

論文誌（査読あり）

- Tsukasa Shiota, Kouki Honda, Kazutaka Shimada, and Takeshi Saitoh. Development and Application of Leader Identification Model Using Multimodal Information in Multi-party Conversations. *International Journal of Asian Language Processing (IJALP)*, 2021. (in press).

国際会議（査読あり）

- Tsukasa Shiota, Kazutaka Shimada, Shinji Nogami, and Shuhei Fukuyama. Related Article Extraction Using Learning to Rank. In *Proceedings of 2020 International Conference on Asian Language Processing (IALP)*, pp. 7-12, 2020.

- Tsukasa Shiota, Kouki Honda, Kazutaka Shimada, and Takeshi Saitoh. Leader Identification Using Multimodal Information in Multi-party Conversations. In *Proceedings of 2020 International Conference on Asian Language Processing (IALP)*, pp. 50-55, 2020.

- Tsukasa Shiota and Kazutaka Shimada. The Discussion Corpus toward Argumentation Quality Assessment in Multi-party Conversation. In *Proceedings of the 9th International Conference on Learning Technologies and Learning Environments (LTLE)*, pp.280-283, 2020.

- Tsukasa Shiota, Takashi Yamamura, and Kazutaka Shimada. Analysis of Facilitators' Behaviors in Multi-party Conversations for Constructing a Digital Facilitator System. In *Proceedings of the 10th International Conference on Collaboration Technologies (CollabTech)*, pp.145-158, 2018.

国内会議（査読なし）

- 塩田 宰, 嶋田 和孝. マルチモーダル情報を用いた複数人議論の品質評価. 人工知能学会 言語・音声理解と対話処理研究会 (SIG-SLUD) 第 91 回研究会, 2021. (to appear).

- 塩田 宰, 嶋田 和孝, 野上 真司, 福山 修平. ランキング学習を用いた関連記事候補の抽出. 電子情報通信学会 言語理解とコミュニケーション研究会 (NLC) 第 16 回テキストアナリティクスシンポジウム, pp. 1-6, 2020.
  NLC 研究会 学生研究賞（Best Student Paper Award）

- 塩田 宰, 嶋田 和孝. 議論参加者の陳述評価に向けた複数人議論コーパスの構築. 電子情報通信学会 言語理解とコミュニケーション研究会 (NLC) ヴァーバル・ノンヴァーバル・コミュニケーション研究会 (VNV) 合同研究会, pp. 1-6, 2020.

- 本多 幸希, 塩田 宰, 嶋田 和孝, 齊藤 剛史. マルチモーダル情報を考慮した議論の取りまとめ役推定. 電子情報通信学会 言語理解とコミュニケーション研究会 (NLC) ヴァーバル・ノンヴァーバル・コミュニケーション研究会 (VNV) 合同研究会, pp. 27-32, 2020.

- 塩田 宰, 山村 崇, 嶋田 和孝. マルチモーダル情報を考慮した議論の取りまとめ役推定. 言語処理学会 第 24 回年次大会 (NLP), pp. 833-836, 2018.

- 塩田 宰, 山村 崇, 嶋田 和孝. 複数人対話における議論の取りまとめ役の推定. 平成 29 年度電子情報通信学会 九州支部 第 25 回学生会講演会, 2017.