対話的議論の自動評価に向けたディベートデータセットの構築

澤田慎太郎 [1]　中川智皓 [1]　新谷篤彦 [1]　井之上直也 [2,3]
[1] 大阪府立大学　[2] 東北大学　[3] 理化学研究所
{szb03072@edu, chihiro@me, shintani@me}.osakafu-u.ac.jp
naoya-i@ecei.tohoku.ac.jp

1 はじめに

会社の会議などの意思決定を有する議論の場において，発言内容を評価し，明快でわかりやすい話の構成や説得力のある内容をフィードバックすることは，発言者の議論能力の向上につながる．しかし，発言内容を一つ一つ評価するのは評定者に負担がかかるため，評価を自動化する必要性がある．ここでの議論の自動評価とは，話者の発言がどれほど聴衆の意思決定に影響を与えるかを，自動的かつ定量的に評価することである．特に，事前の発言との噛み合いを考慮しながら，時系列ごとに発言内容の説得力を評価することが特徴である．本研究では，図1に示すような対話的議論の自動評価について検討する．

議論の自動評価は，議論マイニング (Argument Mining) の分野において盛んに研究が進められている [1, 3, 9, 10]．しかし，これらの研究は主にモノローグな論述文により与えられる議論を対象としているため，本研究が対象とする対話的な議論を十分に評価することができない．対話的な議論では，わかりやすい構成，明示的表現に加えて（図1，青線），前者の主張に対する反論，意見の噛み合い（図1，赤線），などを判定する必要があるため，モノローグな論述文の評価と，対話的な議論の評価の間には研究課題の隔たりが存在する．

そこで本研究では，スコア付きの対話的議論のデータセットの構築を行い，既存の技術の到達点及び今後の課題を検証することにより，対話的な議論の自動評価の研究の足がかりを築く．初期の探索的研究であるため，本質的な課題が浮き上がるデータセットの構築をするべく，対話空間の場としてある程度制約のかかったディベートを研究対象とする．ディベートは自由な議論とは異なり，ルールに基づいて順番に話すため，発話の重なりがない，話し方がフォーマルである，論点ごとにわけて討論する，など，対話的議論の評価課題の本質に集中して取り組めるという利点がある．

2 背景

2.1 関連研究

議論の品質の評価をする上で，様々な指標が存在する [11]．対話的議論においては，Yohan Jo et al. の研究 [2] が示すように，議論を通してスタンスが変わったか否かが重要な指標であると考える．しかし，実際に議論を評価する際，二値分類であると細かな変化が不明であり，適切なフィードバックができない問題がある．他に，Persing et al. [8] のような，説得力を指標とした研究も存在し，これはスタンスの変化と密に関わる指標であると考える．そこで本研究では，聴衆のスタンスを細かく段階付けて，説得力も同時に評価できるようスコア付けを行い，対話的議論の自動評価を検討する．

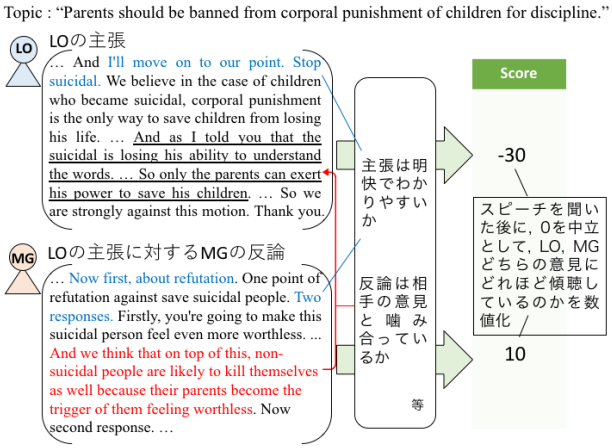

図1: 対話的議論の自動評価課題の全体像．

2.2 パーラメンタリーディベート

パーラメンタリーディベートとは，英国議会を模してゲーム化された議論の形式である．ここでは，肯定側 (Government) 3 名と否定側 (Opposition) 3 名が，与えられた論題に対し，それぞれ 2 つずつ論点を掲げて，即興で聴衆 (Judge) を説得する話し合いの方式を取り扱う．肯定側，否定側のいずれのサイドになるかはスピーカが決めることはできず，論題発表前に主催者によって決定される．論題はディベート開始の 15〜30 分前に発表されるため，話者は短時間で論点をまとめ，聞き手を説得する．

図2に示すように，肯定側の一人目である PM から 3 分間の主張が始まり，次に否定側の一人目である LO が発言する．このように左から右へと議論が進められていき，肯定側の三人目である PMR の主張で試合が終わる流れとなる．スピーカごとに話す内容，順番，時間が決まっており，聴衆は個人的な考えや偏見を含めずに勝敗を決めるといった特徴がある．

パーラメンタリーディベートでは，どのように議論が推移したかをわかりやすく整理するため，スピーカだけでなく，聴衆もフローシート (図2, 下段) を用いることが多い．スピーカがディベートの流れに沿ってスピーチを行えば，フローシートに示される括弧すべての箇所に議論が埋まっていく仕組みである [12]．これら括弧に埋められた各議論は，立論や反論といった

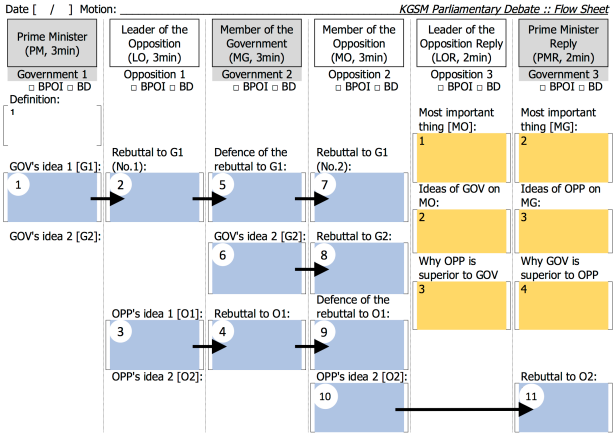

図2: パーラメンタリーディベートの流れおよびフローシート．

ある程度まとまりをもつ主張を指し，本稿では**発言**と呼ぶ．

3　データセット構築

3.1　スコアの付与方法

本研究では，フローシート上のそれぞれの発言に対し，聴衆のスタンス（聴衆が各発言を聞いた時点で肯定側と否定側のどちらを支持しているか）を整数値によりスコア付けした．

スコア付けには，図3に示す議論の強弱可視化シートを用いた．肯定側が掲げる論点をそれぞれ G1, G2, 否定側が掲げる論点を O1, O2 と定義し，各論点ごとにグラフを区別し，議論の勝敗に到るまでの推移を可視化したシートである．シートは大きく分けて，(1) 論題記述欄, (2) 論点記述欄, (3) 各発言のスコア記述欄（青/オレンジによるハイライト部分），の 3 つからなる．

スコア記述欄のグラフの横軸はスピーカ，縦軸はスコアを表す．スコアの範囲は-50~50 の整数値（11 段階の極カテゴリー尺度）であり，負の端が否定側を，正の端が肯定側を支持していることを意味する．例えば，図3における青ハイライト 1 番のスコアは，肯定側の一人目である PM の G1 上の発言 1 に対応する．スコアは発言を聞いた直後の聴衆の肯定側・否定側のどちらに説得されているかを数値化したものであるため，各論点での勝敗結果は PMR のスピーチに対するスコアの正負によって判別できる．最終的には，図3に示される 1~11 の合計 11 個の発言とスコアが対応したデータを一つのディベートから取得する[*1]．

3.2　データ収集の場

一般社団法人パーラメンタリーディベート人財育成協会 (PDA) が主催する以下の 4 つの大会に参加し，データを収集した [7]: (A) PDA 全国高校即興型英語ディベート合宿・大会 2019, (B) 第 3 回 PDA 高校生即興型英語ディベート全国大会 2017, (C) 第 4 回 PDA 高校生即興型英語ディベート全国大会 2018, (D) 第 5 回 PDA 高校生即興型英語ディベート全国大会 2019.

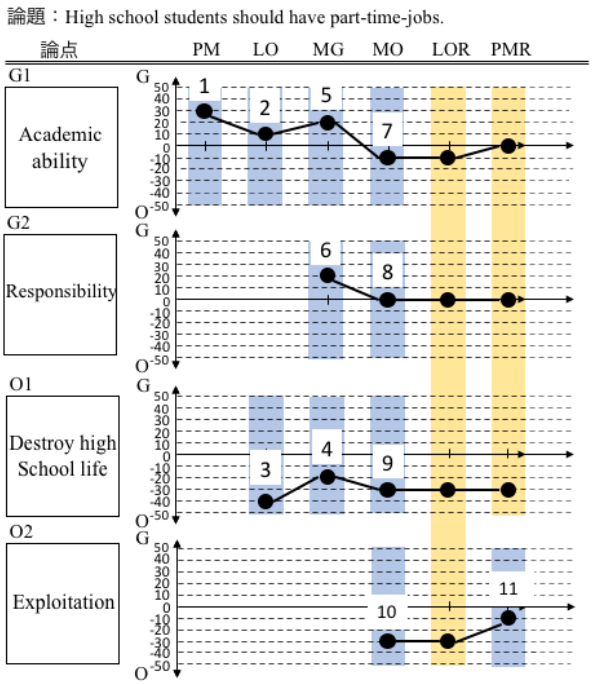

図3: 議論の強弱可視化シート．

(A) では 2 日にわたり, 練習 3 試合, 予選 3 試合, 教員マッチ 1 試合, 準決勝, 決勝と試合が行われた．参加高校数は 28 校であり，生徒, 教員合わせて約 200 名が参加した．(B), (C), (D) ではどの年も 2 日にわたり，予選 4 試合, 準々決勝, 準決勝, 決勝と試合が行われた．参加高校数は 60 校を超え，生徒, 教員合わせて約 270 名が参加した．どの大会も予選試合では，英語が堪能でない生徒もみられたが，準々決勝以降はどの試合も内容，英語の流暢さともに差し障りなく，発言のデータとして質が高いことが言える．

録音機と議論の強弱可視化シートを用いてデータ収集を行った．試合の音声をテキストに書き起こし，1 試合あたり 2 人以上の聴衆からスコアを得た[*2]．各試合ごとに収集したスコアの平均をとり，一つの発言に対して一つのスコアが対応するデータセットを構築した．なお，スコアの平均を取った後，これらを 11 段階のスコアに戻すために，整数値 {-5, ..., 5} のスコアにスケーリングし直している．

3.3　収集結果

統計　収集したデータの論題と発言の数を表1にまとめる．一例として，(A) の教員エキシビションマッチでのスコアの推移と標準誤差を図4に示す．各論点ごとに発言前後のスコアを比較すると，肯定側チームの発言後では正方向に，否定側チームの発言後では負方向にスコアが変動している．これより，各発言は少なからず聴衆の肯定側もしくは否定側への投票行動につながる，説得力のある内容であることがわかる．

アノテータ間の一致率　計算機による自動評価問題としてのデータセットの有用性を検証するため，異なる 2 人以上の聴衆が

[*1] LOR の発言, PMR の G1,G2,O1 の発言は各論点ごとに分けずに内容を総合した発言であるため, 今回は対象外とした．

[*2] スピーチ時間の制限ゆえ, 話されなかった発言箇所がいくつかあり, この場合はスコアを付与していない．

表1: 収集データの論題とディベートの数（カッコ内は発言の数）.

場	試合	論題	ディベート数
(A)	教員マッチ	High school students should have a part-time job	1 (11)
	準決勝, 決勝	People should have a microchip implant in their own bodies	5 (50)
	決勝	Companies should introduce the four-day-workweek system	1 (10)
(B)	予選 3	Grade skipping should be introduced in compulsory education	5 (50)
	予選 4	Government should restrict that time spent on online games	1 (11)
(C)	予選 1	Using private homes as hotels for school trip should be banned	2 (21)
	予選 2	Japan should raise the pension age to 70 years old	4 (38)
(D)	準々決勝	Disaster forecasts should be issued from a single source	3 (33)
	準決勝	Japan should pay the full costs of US military bases in Japan	2 (21)
	決勝	Parents should be banned from corporal punishment of children for discipline	1 (10)

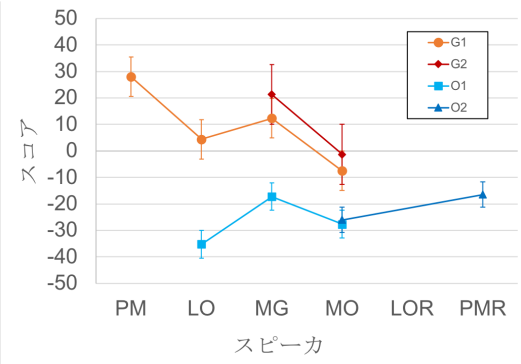

図4: 教員エキシビションマッチでのスコアの推移.

付与したスコアの一致率を計算した．ここでは Krippendorff's alpha を用いた．

収集したデータの中には，聴衆の間で勝敗結果が分かれた試合も存在した．勝敗結果別と全体でそれぞれスコアの一致率を計算したところ，全体よりも勝敗結果別の方が一致率が高くなる傾向がみられた．よって本研究では，聴衆間で勝敗結果が分かれた試合に関しては，多数決にのっとり，勝利したチームを支持していた聴衆のスコアのみを利用しデータセットを構築した．

本実験の学習，評価に使用したデータでのスコアの一致率の平均値は，$\alpha_{\text{average}} = 0.686$ であった (substantial agreement [4])．今回収集したデータの中で，一致率が低い試合は 2 つあったが，これら含めた全体のスコアの一致率の平均値は 0.649 であり，スコア予測問題として成立することが示唆される．

一致率が低い 2 試合のディベートにおいて，どちらも MO の発言で論点をまたいだ反論を確認した．明示的にどの論点に対する反論の発言か主張がなかったため，聴衆がスコア付けする際に困惑したと考えられる．

4 評価実験

本稿のデータセットの難易度を推定するために，基礎的なベースラインモデルを構築しその性能を評価する．

4.1 実験設定

モデル まず，発言の特徴量を獲得するため，事前学習言語モデルである ALBERT[*3] を用いた [5]．ALBERT は，従来の BERT よりも少ないパラメタ数にも関わらず様々な自然言語処理のタスクに対する有効性が示されており，事例数が少ない今回の評価実験に適していると考えられる．

スコアの予測対象となる発言は議論の部分的スピーチであるが，対話的議論ではそれまでに話された内容を総合して評価する必要がある．このため，ALBERT への入力を "[CLS] 対象となる発言 [SEP] 対象となる発言以前の議論 [SEP]" とし，最終的な発言の特徴量として，対象となる発言以前の議論を含む発言全体の特徴量を示す [CLS] の埋め込みを獲得した．

次に，ALBERT により得られた [CLS] 埋め込みに対し，非線形変換を適用し，線形結合層で 1 次元の実数へと回帰する．具体的には，$\boldsymbol{u} \in \mathbb{R}^D$ を [CLS] 埋め込み, $W \in \mathbb{R}^{D \times D}$ を重み行列，$\boldsymbol{w} \in \mathbb{R}^D$ を重みベクトル，b_1, b_2 をバイアス項とおき，次式のような変換を適用する:

$$\boldsymbol{v} = \tanh(W\boldsymbol{u} + b_1)$$
$$y = \tanh(\boldsymbol{w} \cdot \boldsymbol{v} + b_2) \cdot 5$$

ここで得られた出力 y が予測スコアとなる．

学習 損失関数を平均二乗誤差とし，訓練データに対して，AdamW [6] で最適化する．学習率を 1e-6，エポック数を 20 とし，検証データに対して最適なモデルを取得する．この際，ALBERT の fine-tuning も同時に行う．

実験設定 現在の技術でどの程度問題を解くことができるのか検証する．そこで，予測精度の基準値を得るため，ランダムにスコアを予測する場合の性能と比較をした (**random**)．さらに，今回のモデルの評価方法として，(1) テストデータの論題と訓練データの論題が同じ場合 (**indomain 設定**), (2) テストデータの論題が訓練データに含まれない場合 (**outdomain 設定**)，の 2 種類を実施した．実際に議論の評価をする際，その場に応じた論題の学習データが必要であるか否か検証するためである．少数のデータを有効に活用するため，leave-one-out 交差検証法で評価を行った．具体的には，outdomain 設定では，

[*3] https://github.com/huggingface/transformers, albert-base-v2

表2: 各試合における平方平均二乗誤差.

試合	random	indomain	outdomain
(A) 教員マッチ	4.4	-	1.9
(A) 準決勝, 決勝	3.8	1.9 (±0.3)	1.9
(A) 決勝	3.0	-	2.4
(B) 予選 3	4.0	1.9 (±0.5)	1.9
(B) 予選 4	3.7	-	2.3
(C) 予選 1	4.2	-	2.3
(C) 予選 2	3.6	2.1 (±0.6)	2.2
(D) 決勝	4.5	-	1.8
(D) 準々決勝	3.8	2.3 (±0.3)	2.4
(D) 準決勝	3.6	-	2.1
Overall	3.9 (±0.4)	2.0 (±0.5)	2.1 (±0.2)

任意の 1 トピックをテストデータ，任意の 1 トピックを検証データ，それ以外を訓練データとした．indomain 設定では，任意の 1 試合をテストデータ，任意の 1 試合を検証データ，それ以外を訓練データとした[*4]．テストセットにおけるモデルの性能評価は，平方平均二乗誤差による．

4.2 実験結果

実験結果を表2に示す．indomain 設定において，random よりも大きく予測精度が高いことがわかる．これより，今回使用したモデルは発言の品質に関連する何らかの特徴を読み取っていると考えられる．しかし，random より高い精度で予測できたとはいえ，予測誤差は 2 点ほどあり，対話的議論の自動評価の予測精度をあげるには，まだ課題は多いと言える．また，indomain 設定と outdomain 設定における予測精度を比較すると，前者の方が僅かに予測精度が高いが，際立った差は見られなかった．これより，モデルはテスト対象とは異なる論題からも品質予測に関する知識を一般化して学習することができ，未知の論題データに対しても，有効であることが示唆される．同時に，現実的にディベートは異なる論題の試合が多く存在する状況にあり，このような状況下で複数の異なる論題のデータが予測に有効活用できることが示唆される．

5 おわりに

本論文では，対話空間の場としてある程度制約がかかったパーラメンタリーディベートを取り上げて，対話的な議論の研究に取り組むためのスコア付き対話的議論のデータセットを構築した．また，構築したデータセットを用いて自動品質評価の予備実験を実施し，高精度な予測の実現には大きな課題があることを示した．

今後の課題として次の二つがある．まず，今回収集したデータは比較的小規模であり，十分に自動品質評価モデルの性能を評価できたとは言い難いため，さらに事例数を増やしていく予定である．また，自動品質評価モデルについては，1 発言に含まれる単語数が多いため，品質評価に重要な単語を優先して特徴量を抽出するような機構を導入することが考えられる．さらに，各発言の相互作用をより明示的に考慮できるようにするために，各発言の埋め込みを個別に取得し，発言間の噛み合わせを考慮する層を追加することが考えられる．

[*4] 訓練，検証，テストにデータを分割する必要上，3 試合以上ある論題で実験を行った．

謝辞

本研究は，JST 未来社会創造事業 JPMJMI17C7 の助成を受けたものであり，ここに感謝の意を表したい．

参考文献

[1] Ivan Habernal, Patrick Pauli, and Iryna Gurevych. “Adapting serious game for fallacious argumentation to German: pitfalls, insights, and best practices”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[2] Yohan Jo et al. “Attentive Interaction Model: Modeling Changes in View in Argumentation”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 103–116.

[3] Zixuan Ke and Vincent Ng. “Automated Essay Scoring: A Survey of the State of the Art”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 6300–6308.

[4] Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[5] Zhenzhong Lan et al. “ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS”. en. In: (Sept. 2019), p. 17.

[6] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *arXiv:1711.05101 [cs, math]* (Jan. 2019). arXiv: 1711.05101.

[7] *PDA 一般社団法人パーラメンタリーディベート人財育成協会*. ja. https://pdpda.org/.

[8] Isaac Persing and Vincent Ng. “Lightly-Supervised Modeling of Argument Persuasiveness”. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 594–604.

[9] Kaveh Taghipour and Hwee Tou Ng. “A Neural Approach to Automated Essay Scoring”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1882–1891.

[10] Henning Wachsmuth et al. “Argumentation Quality Assessment: Theory vs. Practice”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 250–255.

[11] Henning Wachsmuth et al. “Computational Argumentation Quality Assessment in Natural Language”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 176–187.

[12] 中川 智皓. “パーラメンタリーディベート”. In: システム／制御／情報 63.4 (2019), pp. 170–175.