マルチモーダル情報を用いた複数人議論の品質評価

Multimodal Argumentation Quality Assessment in Multi-party Conversations

塩田 宰[1*]　嶋田 和孝[1]
Tsukasa Shiota[1]　Kazutaka Shimada[1]

[1] 九州工業大学大学院 情報工学府 先端情報工学専攻
[1] Department of Advanced Informatics, Kyushu Institute of Technology

Abstract: Recently, argumentation quality assessment tasks have attracted a lot of attention in NLP. However, there are a few studies to challenge argumentation quality assessment, focusing on synchronous argumentation. In this study, we build a multimodal multi-party argumentation corpus that is annotated argumentation quality scores. We propose some automatic argumentation assessment methods and report the results of each method. We show that information to deal with should be changed based on the target score to assess.

1　はじめに

近年，教育において問題基盤型学習や協調学習がコミュニケーション能力など知識以外の能力（非定形知）を訓練する手段として注目を集めている．教育において非定形知を訓練する方法の 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．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・フィードバックする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を獲得して議論内容の品質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．

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例としては Automated Essay Scoring（AES）[Ke 19] などがある．しかしながら，話し言葉による議論は構成が書き言葉ほど明確ではなく，言い淀みなどを含む言語を取り扱う必要があるなど，書き言葉とは異なり様々な困難さが存在する可能性がある．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 [武川 18]．

そこで，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を実施する．

2　関連研究

対面形式の議論を対象としたコーパスは複数存在する．Zhang ら [Zhang 16] は競技式ディベートをコーパスとして構築し，60%ほどの正解率で勝者を予測が可能であることを報告している．林ら [林 15] は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，採用試験を模したディスカッションを収録したコーパスを構築している．このコーパスには発話の書き起こしや動作・音声データ，およびコミュニケーションスキルに関連する 5 つの項目について各議論参加者を評価した結果が収録されている．Olshefski ら [Olshefski 20] は教室の協調学習の分析を目的として，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．

1 節でも述べた通り，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，議論の強さや構成など様々な評価軸によって自動評価する AES [Ke 19] がある．インタラクションを対象とした研究では人やグループのパフォーマンスを評価する手法を提案する研究は数多く存在している．例

*連絡先：九州工業大学大学院 情報工学府 先端情報工学専攻
〒 820-8502 福岡県 飯塚市 川津 680-4
E-mail: t_shiota@pluto.ai.kyutech.ac.jp

えば，Okada ら [Okada 16] は複数人議論の参加者のコミュニケーション能力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．他にも，Avci ら [Avci 16] や Murray ら [Murray 18] 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．

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．

3 データセットの構築

著者ら [Shiota 20] はこれまでの研究で 4 人一組による議論を収録した複数人対話コーパスを構築している．本コーパスでは 2 人が 1 つのグループとなり，ある命題に対して賛成と反対の立場から討論する対話と，2 つのグループが納得のいく折衷案を導く合意形成を行う対話を収録している．1 つの対話は約 20 分で，5 つのグループによる討論および合意形成合わせて 10 対話を現在収録しており，合計対話時間約 200 分となっている．対話を記録した映像・音声データを基に，0.2 秒以上の無声区間を区切りとする転記単位で書き起こされた発話のテキスト，議論参加者の骨格や顔の情報を OpenPose[1]，OpenFace[2]によって解析した座標点，そして Surfboard[3]によって解析した発話音声特長量といったマルチモーダル情報が抽出，構造化されたコーパスとなっている．10 対話の合計発話数は 7,449 発話である．

一般に，1 つの命題に対して討論/合意形成を行う場合，様々なトピックについて議論することになる．そこで本研究では，採点対象とする議論をトピックセグメンテーションされた対話のセグメント（以降，議論セグメント）とし，議論セグメントに品質スコアを付与したデータセットを構築する．本研究では，各対話に含まれる議論セグメントを獲得するため，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 [Xu 05] を参考にメイントピックの境界をアノテーションし，採点対象となる議論セグメントを 178 件を獲得した．

次に，自然言語による議論の計算論的品質評価に関する理論 [Wachsmuth 17b] を基に評定軸および基準を作成し，評定作業を実施する．Wachsmuth らの定義する分類体系では，議論の品質は「合理性（Re）」および「有効性（Ef）」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．それぞれに対する従属評価軸は「容認性（GA）」「関連性（GR）」「充足性（GS）」，および「信用性（Cr）」「情動性（Em）」「明瞭性（Cl）」「妥当性（Ap）」「順序性（Ar）」となっている．表 1 に先行研究を基に作成した評価基準の概要を示す．

表 1: 議論セグメントの品質の評定軸一覧

軸	概要
Re	議論が GA から GS を満たしているか
GA	議論の内容/進行方法が適切で許容できるか
GR	議論の内容が議題と関連したものであるか
GS	議論の内容が内省されているか
Ef	議論が Cr から Ar を満たしているか
Cr	議論の内容を信頼することができるか
Em	議論に対してオープンマインドになれるか
Cl	議論の内容が明快か
Ap	議論における各話者の言動が建設的か
Ar	論点や根拠，結論の流れが理解しやすいか

各グループに対して第三者 3 名を割り当て，議論セグメントの品質を評定することを依頼した．評定者は議論の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表 1 の評価指標について詳細な説明を行っている資料を基に議論セグメントを評定した．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，M，H，VH（Very High）の中から決定する．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 2 に各軸のクリッペンドルフの α 係数を示す．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，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．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実施し，分析結果を報告している論文 [Wachsmuth 17a] が存在する．この研究においてもクラウドワーカーによる評定結果のクリッペンドルフ α 係数は-0.27〜0.53 と閾値に到達していない．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．しかし，これ

[1]https://github.com/CMU-Perceptual-Computing-Lab/openpose
[2]https://github.com/TadasBaltrusaitis/OpenFace
[3]https://github.com/novoic/surfboard

表 2: 各評定軸のクリッペンドルフ α 係数

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

らの一致率もまた不十分なため，今後は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．なお，本研究で利用しているコーパスと評定スコアは準備が整い次第本研究室のホームページ[4]にて公開予定である．

4 品質評価手法

本節では，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，本研究における品質評価タスクを形式化し，その後 SVM や深層学習を用いた品質評価モデルを提案する．

4.1 品質評価タスクの形式化

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．

$$\boldsymbol{u}_i = [\boldsymbol{sp}_i; \boldsymbol{t}_i; \boldsymbol{b}_i; \boldsymbol{f}_i; \boldsymbol{a}_i] \tag{1}$$

$\boldsymbol{sp}_i$ は 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 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 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 i 番目の発話中の顔情報を表現するベクトル，$\boldsymbol{a}_i$ は 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．

4.2 SVM

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）[Vapnik 13] を用いたモデルを構築する．SVM では系列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．

4.3 LSTM

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を平均値および標準偏差で時間方向に畳み込んでいるため，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 LSTM を用いた品質評価モデルを提案する．

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．

$$\boldsymbol{h}_i = LSTM(\boldsymbol{u}_i, \boldsymbol{h}_{i-1}, \boldsymbol{c}_{i-1}) \tag{2}$$

上記の手順を繰り返し，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 $\boldsymbol{h}_N$ を獲得する．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．

$$\hat{Y}_{dim} = \mathrm{softmax}(\boldsymbol{W}_s \boldsymbol{h}_N + \boldsymbol{b}_s) \tag{3}$$

$\boldsymbol{W}_s$，$\boldsymbol{b}_s$ は学習パラメータ，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．

$$\hat{y}_{dim} = \arg\max_{y_{dim}} \hat{Y}_{dim} \tag{4}$$

4.4 Attention-based LSTM

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．そこで，LSTM に注意機構を導入した先行研究 [Wang 16][Zhou 16] 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．

まず，LSTM を用いた品質評価モデルと同様に，時刻 i の LSTM の状態を算出する．その後，LSTM の各時刻における出力 h_i に対する重み a_i を次式を用いて計算する．

$$m_i = \boldsymbol{\omega}^T \tanh(\boldsymbol{h}_i) \tag{5}$$

$$a_i = \frac{\exp(m_i)}{\sum_{j=1}^{N} \exp(m_j)} \tag{6}$$

[4] http://www.pluto.ai.kyutech.ac.jp/~shimada/resources.html

$\boldsymbol{\omega}^T$ 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．

$$\boldsymbol{r} = \sum_{i=1}^{N} a_i \boldsymbol{h}_i \tag{7}$$

$$\boldsymbol{h}^* = \tanh(\boldsymbol{r}) \tag{8}$$

tanh() は双曲線正接関数である．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．

$$\hat{Y}_{dim} = \mathrm{softmax}(\boldsymbol{W}_s \boldsymbol{h}^* + \boldsymbol{b}_s) \tag{9}$$

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．

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，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（8）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．

$$\boldsymbol{h}^* = \tanh(\boldsymbol{W}_r \boldsymbol{r} + \boldsymbol{W}_{h_N} \boldsymbol{h}_N) \tag{10}$$

$\boldsymbol{W}_r, \boldsymbol{W}_{h_N}$ 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．

4.5 階層型 LSTM

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 $\boldsymbol{t}_i$ は単語の系列情報を考慮していない問題がある．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 [Tran 17] を参考に構築する．

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．

$$\boldsymbol{h}^{Uttr}_{i,j} = LSTM^{Uttr}(\boldsymbol{w}_{i,j}, \boldsymbol{h}^{Uttr}_{i,j-1}, \boldsymbol{c}^{Uttr}_{i,j-1}) \tag{11}$$

$$\boldsymbol{w}_i = \boldsymbol{h}^{Uttr}_{i,M_i} \tag{12}$$

$LSTM^{Uttr}()$ 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，$\boldsymbol{h}^{Uttr}_{i,j}$ は i 番目の発話の時刻 j における発話 LSTM の隠れ層，$\boldsymbol{c}^{Uttr}_{i,j}$ 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．

$$\boldsymbol{u}_i = [\boldsymbol{sp}_i; \boldsymbol{w}_i; \boldsymbol{b}_i; \boldsymbol{f}_i; \boldsymbol{a}_i] \tag{13}$$

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 $\boldsymbol{h}^{Cont}_N$ を獲得する．

$$\boldsymbol{h}^{Cont}_i = LSTM^{Cont}(\boldsymbol{u}_i, \boldsymbol{h}^{Cont}_{i-1}, \boldsymbol{c}^{Cont}_{i-1}) \tag{14}$$

$LSTM^{Cont}()$ は単語ベクトルの系列をエンコードする発話 LSTM，$\boldsymbol{h}^{Cont}_i$ は時刻 i における文脈 LSTM の隠れ層，$\boldsymbol{c}^{Cont}_i$ 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．

$$\hat{Y}_{dim} = \mathrm{softmax}(\boldsymbol{W}_s \boldsymbol{h}^{Cont}_N + \boldsymbol{b}_s) \tag{15}$$

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．

5 実験

本節では，4 節で提案した手法を本研究で構築したデータセットに適用した結果についてまとめる．

5.1 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 の事前学習モデル[5]を用いて獲得する．本研究では [CLS] トークンに対応する BERT の 11 層目を $\boldsymbol{t}_i$，各単語に対応する BERT の 11 層目を $\boldsymbol{w}_{i,j}$ 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とした．発話の音声情報を表すベ

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

表 3: 実験で用いる正解スコアの分布

評価軸	L	M	H
Re	13	89	76
Ef	9	97	72

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とした．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．

本研究では α 係数が相対的に高い主要評価軸の Re と Ef を正解ラベルに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 3 に示す．

本研究における品質推定タスクは 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，深層学習のモデルの損失は 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．最適化手法は SGD [Bottou 91] に慣性項（Momentum） [Qian 99] を付与した SGD Momentum（慣性項 $\alpha = 0.95$）を利用した．ハイパーパラメータについては，隠れ層を 500 次元，バッチサイズを 32，エポック数を 50，学習率を 0.01，ドロップアウトを 0.2，重み減衰を 0.001 に設定した．

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 10 対話交差検証を実施した結果を実験の評価値とする．評価指標には各ラベルについて F 値を算出し，そのマイクロ平均を利用した．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を報告，議論する．

5.2 実験結果

表 4 に，各モダリティの組み合わせによる評価性能を示す．baseline は全てのセグメントの品質を M を予測した場合の評価値であり，太字は各モダリティの組み合わせにおいての最高値を示す．

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．

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，合理性（Re）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向上しないと考えられる．一方，有効性（Ef）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を求めたりする場合にはアイコンタクトやボディランゲージなど言外で何か工夫を行うことが考えられる．そのため，有効性の品質評価の性能は言語・非言語の両方を導入することで向上すると推測できる．

6 おわりに

本研究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と品質ラベルのペアデータセットを作成し，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．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ていない．そのため今後は，更なる数の議論をコーパス化し，データを拡充することが喫緊の課題としてあげられる．また，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ため，チェックリスト方式を導入や評価軸の定義を簡易化するなどの工夫を行い，より信頼性の高い品質ラベルを付与する手続きの構築が必要である．本研究では実験で合理性と有効性それぞれ F 値で 0.459，

表 4: 各モダリティの組み合わせによる各モデルの評価性能

評価軸	モデル	評価性能（マイクロ平均の平均値）							
		T	TB	TF	TA	TBF	TBA	TFA	TBFA
Re	baseline	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333
	SVM	**0.451**	0.338	0.337	0.343	0.333	0.317	0.320	0.340
	LSTM	0.387	**0.398**	0.392	0.380	0.410	0.379	0.388	0.360
	A-LSTM1	0.412	0.392	0.387	0.399	0.371	0.359	**0.398**	0.398
	A-LSTM2	0.371	0.337	0.400	**0.420**	0.404	0.374	0.387	0.339
	HLSTM	0.359	0.354	**0.459**	0.388	**0.415**	**0.391**	0.370	**0.405**
Ef	baseline	0.384	0.384	0.384	0.384	0.384	0.384	0.384	0.384
	SVM	**0.459**	0.382	0.383	0.436	0.384	0.392	0.406	0.379
	LSTM	0.428	**0.478**	**0.438**	**0.476**	**0.467**	**0.472**	**0.486**	0.435
	A-LSTM1	0.433	0.470	0.426	0.468	0.450	0.396	0.444	**0.490**
	A-LSTM2	0.452	0.412	0.416	0.436	0.443	0.425	0.401	0.404
	HLSTM	**0.459**	0.433	0.416	0.379	0.414	0.440	0.431	0.451

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．更なる精度向上を目指し数理議論学 [若木 17] の知見や知識グラフ [Al-Khatib 20] などの世界知識を用いた評価モデルの改良を行っていく必要がある．

謝辞　本研究は科研費 20K12110 の助成を受けたものです．

参考文献

[Al-Khatib 20] Al-Khatib, K., et al.: End-to-End Argumentation Knowledge Graph Construction, in *Proceedings of AAAI*, pp. 7367–7374 (2020)

[Avci 16] Avci, U. and Aran, O.: 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)

[Bottou 91] Bottou, L.: Stochastic Gradient Learning in Neural Networks, in *Proceedings of Neuro-Nîmes 91* (1991)

[Ke 19] Ke, Z. and Ng, V.: Automated Essay Scoring: A Survey of the State of the Art, in *Proceedings of IJCAI*, pp. 6300–6308 (2019)

[Murray 18] Murray, G. and Oertel, C.: Predicting Group Performance in Task-Based Interaction, in *Proceedings of ICMI*, pp. 14–20 (2018)

[Okada 16] Okada, S., et al.: Estimating Communication Skills Using Dialogue Acts and Nonverbal Features in Multiple Discussion Datasets, in *Proceedings of ICMI*, pp. 169–176 (2016)

[Olshefski 20] Olshefski, C., et al.: The Discussion Tracker Corpus of Collaborative Argumentation, in *Proceedings of LREC*, pp. 1033–1043 (2020)

[Qian 99] Qian, N.: On the Momentum Term in Gradient Descent Learning Algorithms, *Neural Networks*, Vol. 12, No. 1, pp. 145 – 151 (1999)

[Shiota 20] Shiota, T. and Shimada, K.: The Discussion Corpus toward Argumentation Quality Assessment in Multi-Party Conversation, in *Proceedings of LTLE*, pp. 280–283 (2020)

[Tran 17] Tran, Q., et al.: A Hierarchical Neural Model for Learning Sequences of Dialogue Acts, in *Proceedings of EACL*, pp. 428–437 (2017)

[Vapnik 13] Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer science & business media (2013)

[Wachsmuth 17a] Wachsmuth, H., et al.: Argumentation Quality Assessment: Theory vs. Practice, in *Proceedings of ACL*, pp. 250–255 (2017)

[Wachsmuth 17b] Wachsmuth, H., et al.: Computational Argumentation Quality Assessment in Natural Language, in *Proceedings of EACL*, pp. 176–187 (2017)

[Wang 16] Wang, Y., et al.: Attention-based LSTM for Aspect-level Sentiment Classification, in *Proceedings of EMNLP*, pp. 606–615 (2016)

[Xu 05] Xu, W., et al.: Coding Instructions for Topic Segmentation of the AMI Meeting Corpus Version 1.1 (2005)

[Zhang 16] Zhang, J., et al.: Conversational Flow in Oxford-style Debates, in *Proceedings of NAACL*, pp. 136–141 (2016)

[Zhou 16] Zhou, P., et al.: Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification, in *Proceedings of ACL*, pp. 207–212 (2016)

[若木 17] 若木 利子, 新田 克己：数理議論学, 東京電機大学出版局 (2017)

[武川 18] 武川 直樹　他：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)

[林 15] 林 佑樹　他：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)