

プロダクトディメンジョンを評価指標とした グループ対話のアウトプット推定

Estimating performance of group interaction based on product dimension

三浦 郷^{*1}
Go Miura

岡田 将吾^{*1}
Shogo Okada

^{*1}北陸先端科学技術大学院大学

School of information Science, Japan Advanced Institute of Science and Technology(JAIST)

This paper focuses on developing a model for estimating the quality of discussion using multimodal features. For this purpose, we use a group meeting corpus including audio signal data of participants observed in 30 meeting sessions. Also, four annotators watch conversation transcripts and annotate the score about quality of discussion using product dimension which is a sociological criteria. We extracted various kinds of features such as spoken utterances, acoustic features, speaking turns. First, binary (high or low) classification models are trained to infer the annotated score from these features using support vector machine. Second, binary classification models are developed to infer the quality of unknown discussion task. Experiments results show that multimodal model archived 0.92 as the classification accuracy and task independent model archived 0.73.

1. はじめに

意思決定を集団で行う理由の一つは、個人の情報処理能力の限界を克服することである。集団意思決定に関して、社会心理学の領域で研究が行われている。[Ueda 97]では、集団の意思決定は相互作用により認知的刺激を受けるなどメリットがある一方、多数派の同調圧力により少数派の自分の意見を表出できない場合のように、必ずしも個人の意思決定より集団の意思決定の方が良い結果を得るとは限らないと報告されている。集団の意思決定方法やプロセスの質を評価することが出来れば、意思決定方法の見直しや、個人の意思決定結果の尊重など、集団意思決定におけるデメリットを緩和できる可能性がある。本研究では、集団意思決定のための方法の一つであるグループディスカッションを題材として、ディスカッションの質を評価し、質の改善・向上を支援するシステムに応用可能な技術の提案を目指している。

本研究はディスカッションを通じて観測できる参加者個人の発言の仕方、聞き方と発言内容から、その個人が属するグループで行われるディスカッションの質を推定するモデルを機械学習により構築し、評価することを目的とする。本研究で対象とするグループディスカッション (Group Discussion: 以降GDと略記する)の質を、[Hackman 67]における評価指標に基づき定義する。この評価指標は書き言葉に対して適応可能であり、本研究ではディスカッションの発話内容を書き起こしたテキストを評定者が閲覧し、[Hackman 67]を元にGDの質を評価している。本研究の貢献は以下の2点である。

1. グループアウトプットの推定モデルの構築

40名の参加者による計9時間を超える会話データ [林 15]から、評定者4名により評価されたGDの質を推定するモデルを学習する。評価実験の結果から、最大で9割以上の精度でGDの質の高低を推定できることを示す。

2. タスクの種類に依存しない推定モデルの構築

40名の参加者は、グループごとに「インバスケケット型」課題1つと、「ケーススタディ型」課題2つの課題に取り

組んでいる。3つの課題から学習する組み合わせを変えて、課題に依存せずにGDの質を推定するモデルの構築を目指す。評価実験の結果から、最大で7割以上の精度で訓練データに用いていないタイプのGDの質の高低を推定できることを示す。

2. 関連研究

[Avci 16]らは、マルチモーダル情報の特徴量に用いて、グループアウトプット (GDの質)を推定するモデルの構築を行った。アウトプットの高・低の2値分類タスクにおいて、Coupled Hidden Markov Model (CHMM)を改良したIM (Influence Model)とSupport Vector Machine (SVM)を用いたモデルが最大の分類精度を示し、91%であった。[Murray 18]らは、マルチモーダル特徴量に言語に関する情報を含めて、グループアウトプットを推定する回帰予測タスクを行った。データサイズを増加させる転移学習を行い、学習器にRandom Forestを用いたモデルが最良の精度を示し、MSE (平均二乗誤差)は64.4であった。

[Avci 16]と[Murray 18]で用いられたELEAコーパス [Sanchez 13]には、Winter survival taskと呼ばれるインバスケケット型課題にグループで取り組んだ際に得られたマルチモーダルデータが記録されている。この課題はロールプレイング形式となっており、サバイバルを想定した時に必要なアイテム群 (水や専門書など)の重要度を議論しランキングする。専門家によるランキング (正解)と、グループで決めたランキングの差を「グループアウトプット (GDの質)」と定義している。一方で、ディスカッションはインバスケケット型課題に限らず様々な種類が存在し、正解が明確に決まっていない場合が多い。上記の正解との差によりパフォーマンスを定義する方法では、正解が明確に存在しないタスクに対して適応ができない。そこで、本研究では課題に依存しないグループパフォーマンスの定義方法を3.1節で提案する。この評価指標を基にアノテーションしたデータセットを用いて、課題に依存しないGDの質の自動推定の精度を検証する。

連絡先: 三浦郷, 北陸先端科学技術大学院大学, 〒923-1211
石川県能美市旭台1丁目1, s1810175@jaist.ac.jp

3. グループディスカッションコーパス

本研究で用いるグループディスカッションコーパス [林 15, Nihei 14] の説明を行う。以降、このコーパスを GD (Group Discussion) コーパスと呼ぶ。参加者 40 名の内、初対面同士の 4 名を 1 グループとして、合計 10 グループを構成している。各グループは 3 つの課題 [林 15] についてディスカッションに取り組み、ディスカッションの過程で得たデータ群をデータセットとして収集した。結果として、10 グループ × 3 つの課題 × 40 名で、延べ 120 名分の参加者のデータセットが GD コーパスに含まれている。GD コーパスに含む情報の内、本研究ではヘッドセットマイクから観測した音声データを用いて後述のマルチモーダル特徴量抽出を行った。

3.1 ディスカッションの質の定義

表 1: プロダクトディメンジョンの評価指標

y1	その課題への回答は、創造性を有していた。
y2	その課題に対して、適切な方向性、適切な議論、適切な回答が行っていた。
y3	製品（議論の履歴）は、過度な困難はなく、理解可能であり、明快である。
y4	製品（議論の履歴）は、何らかの問題に関する観点や立場を取ることが出来ていない、または対処できない程度、適切な粒度で問題を捉えられず、対象となる問題の扱いに失敗している。
y5	製品（議論の履歴）は幾つかの考え方、物、事に難色を示すものである。
y6	製品（議論の履歴）には綿密さや、事例、詳細が欠如している。
y7	製品（議論の履歴）は、具体的に何か行うことに関連した方向性を持っていて、実質的な提案、または建設的な内容である。
y8	製品（議論の履歴）は、今後第三者にフォローされるべき、特定または一般的な行動方針を示したり暗示したりしている。
y9	製品（議論の履歴）は、あるアイデア、目的、または出来事に関する特定の信念や意見を述べたり、強く示唆したりしている。
y10	製品（議論の履歴）は特定の観点を押し通そうとしている。
y11	製品（議論の履歴）に含まれるアイデアはユニークで、新鮮で、あまり見たことがない驚きのあるものである。
y12	製品（議論の履歴）の一般的な見通し、態度、調子、方向性は「否定的」ではなく「肯定的」である。
y13	製品（議論の履歴）は断片化していたり、相互の発言に関連性がなかったりして、調和のとれた首尾一貫している状態と真逆である。
y14	製品（議論の履歴）はユニークで、普通ではなく、珍しい。
y15	製品（議論の履歴）は幾つかの考え方、物、事をサポートしている。
y16	製品（議論の履歴）は独りよがりであり、活性化していない。
y17	製品（議論の履歴）は普通で、ありふれていて、よく見るものである。
y18	製品（議論の履歴）内容が簡潔すぎて短い。
y19	製品（議論の履歴）のプレゼンテーションは適切な方法で行われ、調和がとれている。

本研究では課題に依存しないグループパフォーマンスの指標として、Hackman により定義された Product dimension を用いる。表 1 にその評価指標を示す。プロダクトディメンジョンの指標は書き言葉に対して適応可能であり、本研究においては GD における参加者の発話内容を書き起こしたテキストデータを 4 人の評定者が閲覧し、この評価指標に基づき各グループに評定値を付与した。この 4 人に付与された評定値の合計を各グループが行った GD の質（機械学習における教師ラベル）として定義する。以降、19 個の評価指標をそれぞれ y1, ..., y19 と略記する。評定値のスケールは、y1 が 1 点から 5 点の 5 段階、y2 から y19 は 1 点から 7 点の 7 段階で評価した。ディスカッションのサンプル数は 30 であったが、そのうち 1 つのサンプルに不備があったため、29 のサンプルを実験に用いた。

3.2 アノテーションの一致率の分析

各グループによる GD で観測された発話内容を書き起こし、4 人の評定者がその文章を閲覧して GD の質の評定を行う。評価基準として前述のプロダクトディメンジョンの指標を用いる。

各評定者による評定の一致度を計算し、各評価項目についての信頼性を確認した。一致度の計算に関して、信頼性分析の項目を評定者とし、各項目ごとにクロンバック α 値とクリップンドルフ α 値を算出した。ある項目に対してクロンバック α が 0.7 以上、クリップンドルフ α が 0.667 以上であれば各評定者によるグループに対しての評定値は概ね一致しており、その項目の評定値は信頼性が高いと言える。

表 2 は各評価指標におけるクロンバック α 、クリップンドルフ α を示している。各値が正の場合は四捨五入、負の場合は絶対値を四捨五入し、マイナスをつけたものを記載している。各評価指標の値は閾値より低く、十分な一致度は得られなかった。GD の質のアノテーションは主観的であるため、高い一致率が得られなかったと考えられる。この理由の考察は今後の課題とする。

4. マルチモーダル特徴量の抽出

評定者は GD を書き起こした文章を閲覧して評価しているため、GD の質は主に書き起こした発話内容に関連すると考え、発話内容の言語特徴量を抽出する。また、議論中の参加者の非言語特徴量も GD の質に影響を与えると考え、[岡田 16] を参考に発話ターン特徴量、発話中の韻律特徴量を抽出する。

4.1 発話ターン特徴量

合計発話長：発話区間から算出した発話時間を発話長とし、1 セッション単位（各課題の開始時間から終了時間まで）で発話長の合計を計算する。

合計発話回数：一つの発話区間を発話断片と定義する。発話断片の回数を発話回数とし、1 セッション単位で発話回数の合計を計算する。

合計発話長（1 秒以上）：1 秒以上の発話断片を抽出し、その長さの総和を計算する。

合計発話回数（1 秒以上）：1 秒以上の発話断片を抽出し、その発話回数を計算する。

4.2 韻律特徴量

各発話断片における韻律情報の特徴量を抽出する。4.1 節で得た発話区間情報から、各発話断片の特徴量を抽出する。音声情報の信号処理と韻律特徴量抽出には Praat^{*1} を用いる。

最大ピッチ、最小ピッチ：各発話断片の最大ピッチ、最小ピッチを計算し、1 セッション中の全発話断片に対して平均値を計算する。

ピッチ平均：各発話断片に関して、発話区間でピッチの平均値を計算し、1 セッション中の全発話断片に対して平均を計算する。

最大インテンシティ、最小インテンシティ：各発話断片の最大、最小音圧（インテンシティ）を計算し、1 セッション中の全発話断片に対して平均値を計算する。

音圧の幅：最大、最小インテンシティの差を計算する。

抑揚：最大ピッチと最小ピッチの差を計算する。

話速：発話のシラブル（音節）数を各発話長で除した値を計算する。

4.3 発話内容の言語特徴量

名詞数、動詞数、感動詞数、フィラー数：1 セッション中の発話に含まれる各品詞：名詞、動詞、感動詞の単語の出現回数、フィラーの回数を計算する。

新規名詞数：4.1 節で得た発話区間情報を利用し、各名詞が初

*1 <http://www.fon.hum.uva.nl/praat/>

表 2: 各評価指標のクロンバック α , クリッペンドルフ α

評価指標	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11	y12	y13	y14	y15	y16	y17	y18	y19
クロンバック α	0.68	0.58	0.68	0.43	-0.03	0.64	0.64	0.34	0.67	0.50	0.60	0.78	0.52	0.63	0.34	0.52	0.62	0.54	0.51
クリッペンドルフ α	0.34	0.26	0.29	0.16	-0.01	0.30	0.30	0.08	0.32	0.20	0.27	0.43	0.19	0.28	0.12	0.15	0.27	0.16	0.20

表 3: 各特徴量セットを用いた場合の推定精度の結果. 太字は正解率 0.7 以上を示す.

	評価指標									
	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10
A	0.52	0.27	0.52	0.32	0.48	0.48	0.56	0.68	0.36	0.48
S	0.36	0.50	0.48	0.18	0.52	0.80	0.56	0.76	0.61	0.74
L	0.76	0.00	0.56	0.00	0.56	0.72	0.44	0.68	0.61	0.67
A + S	0.48	0.23	0.48	0.18	0.56	0.76	0.60	0.88	0.14	0.33
A + L	0.32	0.00	0.48	0.00	0.56	0.56	0.68	0.92	0.39	0.22
S + L	0.48	0.12	0.56	0.00	0.56	0.72	0.64	0.72	0.64	0.59
A + S + L	0.24	0.00	0.52	0.18	0.56	0.76	0.76	0.92	0.04	0.30

	評価指標								
	y11	y12	y13	y14	y15	y16	y17	y18	y19
A	0.52	0.54	0.44	0.23	0.48	0.70	0.44	0.26	0.42
S	0.59	0.50	0.52	0.58	0.36	0.29	0.56	0.67	0.35
L	0.70	0.54	0.56	0.61	0.24	0.00	0.68	0.70	0.42
A + S	0.63	0.63	0.52	0.61	0.36	0.56	0.56	0.30	0.35
A + L	0.59	0.50	0.40	0.61	0.28	0.59	0.60	0.41	0.42
S + L	0.78	0.50	0.56	0.46	0.20	0.26	0.52	0.70	0.31
A + S + L	0.59	0.63	0.52	0.65	0.20	0.37	0.60	0.41	0.42

めて発言された時間, 発言した参加者を特定し, 参加者ごとに数え上げた回数を新規名詞数とする. 新規名詞は, 他の参加者に対して新しい提案をする時に使用される傾向がある.

4.4 グループ特徴量への変換

発話ターン, 韻律, 言語の個人特徴量をグループ特徴量に変換する. 4人の参加者の特徴量の最大値, 最小値, 最大値と最小値の差分, 平均値, 標準偏差, 合計値の計6つを計算して各グループの特徴量とする. 変換前の個人の特徴量は前述した17個であり, グループ特徴量に変換後はそれぞれ6つずつのグループ値を計算するため計102個となった.

5. 評価実験

中央値を基準にグループの評価値を高群・低群に分類し, 2クラス分類タスクを行い推定精度を評価する. 評価実験は実験1と実験2の2つを行う. 実験1では, 学習された分類モデルのマルチモーダル特徴量の組み合わせと推定精度を分析することで, GDの質を推定するために有効な特徴量を明らかにする. 実験2では, 3つの問題解決課題ごとに分けた特徴量の組み合わせと推定精度を分析することで, 未知の課題に対するモデルの推定精度を評価する.

5.1 実験手順

評価には1つのGDデータを抜いた交差検定法を使用した. あるグループ*i*から得られたデータをテスト, それ以外の28グループから得られたデータを訓練に用いて分類モデルの評価実験を行った.

二値分類タスクの準備として, 評定値を二値に変換する際,

評定値のうち中央値より高いものを高群, 低いものを低群とし, 中央値にあたる1つのデータサンプル(1グループ)は取り除いた. 分類モデルの推定性能には全テストデータに対する正解率の平均を用いた. 分類学習には, RBFカーネルを用いたサポートベクトルマシン(Support Vector Machine: SVM)を用いた. RBFカーネルのバンド幅 γ を $[1/120, 1/119, 0.01, 0.1, 1, 5, 10]$, SVMにおける損失とマージンの大きさの間のトレードオフを調整するパラメータである C を $[0.01, 0.1, 1, 5, 10]$ の範囲で探索し, 最適パラメータをテストに用いた.

各モダリティの推定精度への寄与を検証するため, 単一モダリティとマルチモーダル特徴量セットを計7種類準備し, 推定精度の比較を行う. 「韻律」, 「発話ターン」, 「言語」の特徴量セットをそれぞれA, S, Lとする. また, マルチモーダル特徴量セットとして, 「韻律と発話ターン」をA + S, 「韻律と言語」をA + L, 「発話ターンと言語」をS + L, 「韻律と発話ターンと言語」をA + S + Lと表す.

5.2 2種類の分類タスク実験

実験1: 各種モダリティを用いた推定精度の評価

GDの質を推定するために有効な特徴量セットを明らかにするため, 特徴量セットの組を変えてモデルを構築し, 推定精度を比較した. 特徴量セットは5.1節で記述した7つの組を用いた.

実験2: 未知の課題に対する推定精度の評価

課題の種類に依存しない, 未知の課題に対する推定精度を評価するための実験を行う. ディスカッションデータ全29サンプルの内, 課題1は9サンプル, 課題2, 3は10サンプルずつ保有している. 課題1, 2, 3の内, 2つの課題を訓練データに, 残り1つの課題をテストデータに用いる. テストデータに使用する課題のデータ10サンプル(課題1は9サンプル)について, 1サンプルずつテストに用いて推定し, 10サンプルの推定精度の平均を算出した.

6. 評価実験結果

実験1の各種モダリティごとの推定結果を表3に示す. 表3では, 各評価指標において, 単一モダリティとマルチモーダル特徴量セット7種類を用いて推定した時の正解率を示している. 実験2の未知の課題に対する推定結果を表4に示す. 表4では, 各評価指標において, 訓練に用いていない各課題のデータをテストに用いた場合の正解率と, その正解率を平均した結果を示している.

実験1の結果:

各種モダリティごとの推定結果について, 評価指標y8「行動方針の示唆」における「韻律 + 言語(A + L)」, 「韻律 + 発話ターン + 言語(A + S + L)」のモデルの推定精度が最大となり, 0.92であった. 評価指標y8において単一モダリティよりマルチモーダル特徴量を用いたモデルが有効であることが示された. また, 推定精度が0.7以上であるモデルが存在したのは, 評価指標y1「創造性」, y6「綿密さの欠如」, y7「建設的な内容」, y8「行動方針の示唆」, y10「多面的な観点の欠

表 4: 各評価指標における推定精度の結果. 太字は正解率が 0.7 以上であることを示す. 各指標において, 0.5 を超える正解率が存在しないものは割愛した.

	テストデータ			
	課題 1	課題 2	課題 3	平均
y1: 創造性	0.39	0.0	0.61	0.33
y6: 綿密さの欠如	0.60	0.41	0.24	0.40
y7: 建設的な内容	0.70	0.0	0.06	0.25
y8: 行動方針の示唆	0.90	0.35	0.94	0.73
y11: 驚くべき内容	0.61	0.47	0.67	0.58
y12: 肯定的な態度	0.70	0.00	0.00	0.23
y14: 独自性	0.85	0.33	0.44	0.54
y15: 意見への支持	0.00	0.81	0.00	0.27
y16: 独りよがりな内容	0.55	0.00	0.00	0.18
y17: ありふれた内容	0.65	0.44	0.38	0.49
y18: 情報の不足	0.39	0.89	0.31	0.53
y19: プレゼンの適切性	0.39	0.33	0.71	0.48

如, y11「驚くべき内容」, y16「独りよがりな内容」, y18「情報の不足」であった. また, 全 19 個の指標のうち 15 個の指標で, 2 クラス分類タスクのチャンスレベル 0.5 を上回るモデルが存在した.

実験 2 の結果:

未知の課題に対する推定結果について, 実験 1 と同様に評価指標 y8「行動方針の示唆」において精度が最大となり, 0.73 であった. 評価指標 y8 に関して, タスクに依存しないモデルを構築できることが示唆された. また, 平均する前の課題別の正解率について, 2 クラス分類のチャンスレベル 0.5 を上回ったのは, 表 4 に示した 12 個の評価指標であった. その他の指標については精度が 0.5 に満たず, 十分な精度には至らなかった. 推定精度を上げるために今後検討を行う.

7. 考察

先行研究である [Avci 16] では, GD の質のレベルの高低を最大 0.91 の精度で分類可能であることが示された. コーパス参加者や使用言語, 評価基準が異なるため, 本研究と上記の先行研究の結果とを直接比較することはできないが, 評定値の 2 クラス分類タスクとして同様に扱えば, 最大 0.92 と同程度の精度でモデル構築可能であることが示された. しかし, 正解率 0.92 を示したのは評定の一致度について信頼性が高くなかった評価指標であり, 考慮に入れる必要がある.

実験 1 の結果から, プロダクトディメンジョンの評価指標 y1「創造性」, y6「綿密さの欠如」, y7「建設的な内容」, y8「行動方針の示唆」, y10「多面的な観点の欠如」, y11「驚くべき内容」, y16「独りよがりな内容」, y18「情報の不足」において, 各種モダリティを用いた推定が可能であることが示された. すなわち, 議論の質をプロダクトディメンジョンの評価指標に基づき定義すれば, 「ケーススタディ型」課題などの正解が明確に決まっていない議論の質を特定の観点からは推定可能であることが期待できる. 実験 2 の結果から, 評価指標 y8「行動方針の示唆」においては課題に依存しないモデル構築が可能であることが示された. しかし, テストデータに用いた課題によって推定精度にばらつきがあること, 評価指標 y8 以外ではモデルの精度が芳しくないことは今後の課題である.

今回行ったアノテーションは, 4 人の評定者の主観に依るところが大きいため, 評定の一致度が低くなったと考えられる.

今後, 4 人の評定者の中で誰の評定が大きく異なっているか詳細を分析し, アノテーションの一致率が高い評定者を選定して信頼性の高いモデルを構築することを目指す. また, 評定者の人数を増やすと評定の一致率がどう変化するか傾向を分析することも今後の課題である.

8. 結論

本研究では, グループ対話の参加者から得られるマルチモーダル情報からディスカッションの質を推定するモデルの構築・評価を行った. 発話ターン, 韻律, 言語といった情報をマルチモーダル特徴量として抽出し, プロダクトディメンジョンの評価指標に基づき評定された各グループの GD の質を推定するモデルを機械学習により構築した. 評価実験の結果, 評価指標 y8「行動方針の示唆」に関して, 高群・低群の 2 クラス分類タスクで最大 0.92 の精度を得た. また, 本実験で対象とした 3 つの課題の内, 2 つの課題のデータを学習したモデルを用いて, 学習に用いなかった 1 つの課題のデータに対して推定した精度の平均は, 評価指標 y8 において最大精度の 0.73 を得た.

参考文献

- [Avci 16] Umut Avci, et al. "Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction." in IEEE Transactions on Multimedia, vol. 18, no. 4, pp. 643-658 (2016)
- [Murray 18] Gabriel Murray, et al. "Predicting Group Performance in Task-Based Interaction" In Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 14-20, ACM (2018)
- [Hackman 67] Hackman, J. Richard, et al. "A set of dimensions for describing the general properties of group-generated written passages" Psychological Bulletin vol. 67, no. 6, pp. 379-390 (1967)
- [Sanchez 13] Sanchez-Cortes, Dairazalia, et al. "Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition" Journal on Multimodal User Interfaces, vol. 7, no. 1-2, pp. 39-53 (2013)
- [林 15] 林佑樹, 他 "グループディスカッション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)
- [岡田 16] 岡田将吾, 他 "マルチモーダル情報に基づくグループ会話におけるコミュニケーション能力の推定" 人工知能学会論文誌, vol. 31, no. 6, pp. AI30-E_1-12 (2016)
- [Nihei 14] Fumio Nihei, et al. "Predicting Influential Statements in Group Discussions using Speech and Head Motion Information" In Proceedings of the 16th ACM International Conference on Multimodal Interaction, pp. 136-143, ACM (2014)
- [Ueda 97] 上田泰, "個人と集団の意思決定-人間の情報処理と判断ヒューリスティックス-" 文眞堂 (1997)