複数人談話における
言語情報と非言語情報を利用した盛り上がり判定

横山 貴彦　　嶋田 和孝　　遠藤 勉
九州工業大学大学院 情報工学府
{t_yokoyama, shimada, endo}@pluto.ai.kyutech.ac.jp

1　はじめに

近年，情報推薦のための人間の関心推定システムや，人間と良好な関係を結ぶことを目的とした対話エージェントを開発するために，人間の対話における盛り上がり箇所を分析・検出する研究がなされている．対話の盛り上がり箇所の自動判定を行った既存研究は，言語情報を扱う研究と，非言語情報を扱う研究に分けられる．言語情報の研究として稲葉らは，2者間のテキスト対話に対して，単語の共起情報に基いた判定を行った [1]．一方，非言語情報の研究として，Gatica-Perez らは，複数人会議対話を対象に，会議中の動画と音声より人の頭の動きや声の高さを特徴量とし，判定を行った [2]．また，豊田らは，2者間の会話を対象に，発話状態 (各話者の単独発話，同時発話，無音) の時間長を用いて，盛り上がりを含む対話雰囲気の自動推定を行った [3]．

これらの研究は言語情報，もしくは非言語情報のどちらか一方のみの情報しか用いていない．これに対して，人間が対話の盛り上がりを知覚する際には，発話内容の意味的な文脈や，話し手の声のトーンなど，言語情報と非言語情報を総合して判断している．このため，実際に対話エージェントなどが対話の盛り上がりを推定する上でも，人間と同じように言語と非言語の両方の情報を吟味して，判定がなされるべきであると考えられる．そこで，本稿では対話の盛り上がり箇所の判定を，言語情報と非言語情報の両方を用いて行い，その有効性を検証する．

2　盛り上がり判定手法

2.1　概要

本稿では，複数人談話を対象に，発話を盛り上がり発話と非盛り上がり発話に判定する手法を提案する．この手法は，盛り上がり判定を行う対話エージェントなどのシステムが，室内の離れた位置から談話を傍聴している状況で用いることを想定している．本稿における発話とは，話者の全ての発言に対して，その話者の沈黙で切り離された発言の系列と定義する．この発話の情報はそれぞれ，話者，発話開始と発話終了の時刻，テキストに起こされた発話内容，収録された音声波形からなる．そして，盛り上がりは,「話者がある発話を行った時点で，その話者が談話にどれだけ積極的か」と定義する．話者の積極性は，第3者が収録された音声より知覚できる，発話内容の意味や声のトーンなどを踏まえて判断するものとする．

本稿では，対話の盛り上がりの自動判定を，全発話に対する盛り上がり発話と非盛り上がり発話の分類問題として取り扱う．この分類は機械学習に基づいて行うものとし，学習する素性には言語情報と非言語情報の両方を利用する．言語情報としては，主に関連語に基づく語彙結束性の素性を，非言語情報としては音声の高さや速さ，発話時間などに基づく素性を用いる．これらの情報が盛り上がり判定において相互補完的に作用し，その精度を向上させることを図る．

2.2　言語情報の素性

機械学習に用いる言語情報の素性として次の3種類の素性を用いる．

- (A)　発話の語彙に関する出現の有無
- (B)　発話の長さ
- (C)　発話の語彙結束性

2.2.1　発話の語彙に関する出現の有無

発話の語彙の出現の有無に関する素性は，発話に特定の語彙が出現していれば盛り上がっている可能性が高いと判断するための素性である．具体的には次の語の有無を用いる．

- A_1:　学習データに3回以上出現した単語 unigram
- A_2:　学習データに3回以上出現した単語 bigram
- A_3:　学習データに5回以上出現した単語の語義

$A_{1,2}$ については，自然言語処理の分野で一般的な素性である．A_3 の単語の語義とは，本稿では『岩波日本語表現辞典』に登録されている 4 万 7000 語に割り当てられた分類情報とする [4]．例えば「テスト」や「再試」という単語には「試験」という語義が割り当てられている．このような特定の語義が盛り上がりに関連する可能性を考慮し，語義の出現の有無を素性の 1 つとして取り入れる．

2.2.2 発話の長さ

発話の長さの素性として，次の素性を用いる．

B_1: その発話の相対的な形態素数 (10 段階)

B_1 における 10 段階とは，全発話を形態素数の多い順から 10 段階に分割し，上位何段階目の発話群に属するか[1]ということである．発話の長さと盛り上がりの関連としては，長い発話を行った話者は多くの情報を伝達したい状態にあると考えられ，対話に積極的である可能性が高いと考えられる．一方，短い発話を行った話者は，相手の発言に対する賛同や驚きなど，反応や印象を優先して伝達したい状態にあると考えられる．このため，発話の短さも盛り上がりに関連している可能性がある．発話の長さと短さの両方に着目するため，B_1 の情報を用いる．

2.2.3 発話の語彙結束性

語彙結束性は，発話間における意味の繋がりの強さを表す素性である．対話において，発話と発話の意味的な繋がりが強いということは，ある話題についての詳細なやりとりがなされている考えられる．そして，話題について詳細な話に踏む込むということは，盛り上がっている可能性が高いと判断できる．本稿では，具体的に次の素性を用いる．

$C_{1,2,3}$: 1,2,3 つ前の各発話との語彙結束性の強さ

C_4: 直前 5 秒間の発話との語彙連鎖数の合計

語彙連鎖は，発話の語彙結束性を調べるために使用されるもので，例えば発話間における共起語対，関連語対の数の合計である．本稿における語彙連鎖数の計算には，発話間における名詞について，同じ語義の単語の数と関連語の数の合計を使用する．関連語には，出現した各名詞について Yahoo!API によって取得される 100 件[2]の関連語を利用する．例えば「SNS」という単語では「twitter」や「アカウント」といった語を取得できる．

[1]事前実験において，発話の形態素数そのものをそのまま素性として加えた場合，分類の精度を著しく低下させたため，形態素数そのものを素性として扱わない．

[2]100 件のうちの多くは時事的な著名人や検索される web サイト名などである．これらの語は，談話に出現する可能性は低く，語彙結束性の判定に悪影響をほとんど与えないと考えられる．

$C_{1,2,3}$ は語彙連鎖数が 4 以上であれば「強い結束性」，1〜3 であれば「弱い結束性」，なければ「結束性無し」，発話が存在せず計算できない場合は，「該当発話無し」の 4 値を持つ [1]．C_4 の素性は，$C_{1,2,3}$ よりも長い文脈を考慮した素性である．本稿が対象とする複数人談話では，やりとりが密になれば 1 秒間に 2 回以上の発話がなされていた．対話では，人間同士が意思伝達や情報交換を行っていることから，その文脈は，数秒以上の単位で意味的なまとまりを持つことが考えられる．このため，最大 3 発話の短い文脈における語彙結束性のみに注目するのは，十分ではない．これを踏まえて，本稿では発話の直前 5 秒間になされた発話を対象に語彙結束性の強さを考慮する．

2.3 非言語情報

言語情報以外に基づく情報として，声の高さや速さなどの 1 つの発話に音声として表出する素性と，時間に注目した発話の密さやタイミングなどの，談話雰囲気を捉える素性を用いる．

(D) 音声として表出する素性

(E) 発話の密集度に基づく素性

(F) 発話のタイミングに基づく素性

2.3.1 音声として表出する素性

人が盛り上がっている状態にある場合は，感情が高揚しているため，非盛り上がり時と比べて発声に変化が現れると考えられる．このため，音声の高さと速さの情報[3]を用いる．声の高さは音声波形の基本周波数，発話の速さは単位時間当たりのモーラ数とし，具体的には次の素性を用いる．

D_1: 発話のその話者における高さ (10 段階)

D_2: 発話のその話者における速さ (10 段階)

発話の長さの素性 B_1 と同様に，10 段階に音声の高さと速さを分けて，盛り上がりの傾向を学習する．ただし，声の高さや速さは話者に依存した情報であるので，高さや速さがその話者において何段階目に属するかとする．

2.3.2 発話の密集度に基づく素性

この素性は，談話のうち発話が密集している箇所は，話者が対話に積極的となっていると考えられるため，盛り上がっている可能性が高いという考えに基づく素性である．具体的には次の素性を用いる．

[3]これ以外の情報として，音声の大きさについても，盛り上がり判定に有効な情報の 1 つとされるが，対話エージェントなどが実際に談話の盛り上がり判定を行う場面を想定した場合，談話参加者の位置や向きによって声の大きさが著しく異なるため，本稿では用いなかった．

E_1: 直前 15 秒間の発話数
E_2: 直前 15 秒間の発話の形態素数
E_3: 直前 5 秒間の発話数のバースト値
E_4: 直前 5 秒間の発話の形態素数のバースト値
E_5: 直前 15 秒間のその話者の発話数
E_6: 直前 15 秒間のその話者の発話数の順位

$E_{3,4}$ におけるバースト値とは発話回数もしくは発話された形態素数の急激な増加を表す指標である．バースト値が大きい場合は，談話が活性化していると言えるため，発話が盛り上がっている可能性が高いと判断できる．バースト値は次の式によって計算する [5]．本稿では，注目区間を発話がなされる直前の 5 秒間とし，事前実験から，X=2，Y=10 に設定した．

$$B = \frac{N}{\sqrt{\bar{A}}} \frac{N-A}{N+A} \tag{1}$$

N: 注目区間における発話数もしくは形態素数
A: 直前 X 区間における発話数もしくは形態素数
$\bar{A}$: 直前 Y 区間における発話数もしくは形態素数

$E_{5,6}$ の素性は，対話の主導権に関する素性である．この素性は，対話の主導権を握っている話者は対話に積極的であるため，その話者の発話は盛り上がっている可能性が高いという考えに基づいている．

2.3.3 発話のタイミングに基づく素性

発話間のタイミングは，盛り上がりと高い関連性を持つ．例えば，発話間の間隔が長く，場が沈黙で満たされている時間が長いほど，話者達は対話に消極的であり，対話は盛り上がっていない状態にあると考えられる．また，話者の発話中に，別の話者が発話するオーバーラップ (発話被り) の発生があれば，その話者は対話に積極的であると考えられるため，対話は盛り上がっている状態にあると考えられる．具体的には，次の素性を用いる．

$F_{1,2,3}$: 発話前の沈黙時間の有無 (0.5,1.0,1.5 秒)
F_4: 発話前の沈黙時間の長さ (秒)
$F_{5,6}$: オーバーラップ発話かどうか
$F_{7,8}$: 被オーバーラップ発話かどうか

$F_{5,6,7,8}$ については，オーバーラップの発生があいづち発話によるものを含む $F_{5,7}$ と含まない $F_{6,8}$ の，それぞれ 2 つの素性を用いる．これは，頻繁発話されるあいづちによって発生したオーバーラップと，それ以外によって発生したもので，盛り上がりとの関連性が異なると考えられるためである．特に，あいづち以外によるオーバーラップ発話は，相手の発話を遮ってまで積極的に意見を述べていると考えることができるため，盛り上がり発話に関連している可能性が高く，判定への寄与度が大きいと考えられる．発話のあいづち判定は，文字数が 7 文字以下で，かつ「はい」「へぇ」などの特定の表層語を含む発話を，あいづちと判定するようにした．

3 盛り上がり判定の実験

3.1 対話コーパスの作成

実験用の対話コーパスの作成は，2 名のアノテータによって行った．まず，提案手法の評価実験に用いる対話データを作成するため，談話の音声を次の条件で 10 セット収集した．

1. 10 名の大学生から 4 名の談話参加者を選択
2. 全ての話者は互いに面識有り
3. 5 分間の談話
4. 収録直前に話題を告知 (「最近の映画」など)
5. 談話は図 1 の個室で行う

次に，機械学習の素性を生成するための，発話への情報の付与を行った．具体的には，発話ごとに発話文の書き起こし，話者情報，時間情報の付与，音声波形の発話ごとの分割である．そして，発話への盛り上がり・非盛り上がりラベルの付与を行った．盛り上がりラベルの付与においては，まず，各アノテータが発話に 1〜5 の 5 段階の盛り上がり度を付与した．このとき，盛り上がり発話に対しては盛り上がり度 4 以上付けるように指示した．そして，各アノテータの情報を統合し，ラベルを付与する際には，盛り上がり度の平均値が 3.5 以上のものを盛り上がり発話とした．このアノテーションの結果，盛り上がり発話は 617 件，非盛り上がり発話は 1099 件であり，2 名のアノテータ間で盛り上がり度 4 以上で一致した盛り上がり発話はそのうちの 212 件 (34.5%) であった．

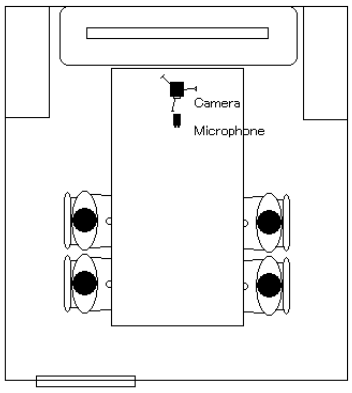

図 1: 対話の収録環境

3.2 実験設定

作成した対話コーパスから，各発話に素性 A〜F の情報を付与したテストセットを作成し，盛り上がり・非盛り上がりラベル分類を機械学習によって行った．機械学習には NaiveBayes を用い，評価には 10 分割交差検証を用いた．評価尺度には，盛り上がり発話に注目した適合率 Pr，再現率 Rc，F 値を用いた．

$$Pr = \frac{\text{盛り上がり発話に判定し，正解した数}}{\text{盛り上がり発話に判定した数}} \quad (2)$$

$$Rc = \frac{\text{盛り上がり発話に判定し，正解した数}}{\text{盛り上がり発話の総数}} \quad (3)$$

$$F = \frac{2PrRc}{(Pr + Rc)} \quad (4)$$

3.3 結果と考察

表 1 に使用した情報に対する盛り上がり判定の結果を示す．表 1 より，言語情報と非言語情報の両情報を用いた判定の結果は，一方の情報のみの結果と比べ，適合率と F 値が良いことが分かった．また，言語情報の素性を用いた場合は適合率が，非言語情報の素性を用いた場合は再現率が高いことが分かった．

表 1: 盛り上がり判定の結果

素性	Pr	Rc	F
言語情報のみ	0.505	0.459	0.481
非言語情報のみ	0.478	**0.596**	0.531
言語情報+非言語情報	**0.535**	0.545	**0.540**

次に，表 2 に適合率，再現率，F 値のそれぞれを最大にする素性の組み合わせの結果を示す．

表 2: 素性の組み合わせと結果

素性	Pr	Rc	F
A, E, F	**0.558**	0.452	0.500
B, C, D, E, F	0.465	**0.627**	**0.545**

最も高い適合率が得られたのは，A, E, F の組み合わせであった．E における発話の密集度の低さや，F の発話間隔の長さは，それぞれ非盛り上がり箇所を推定するのに効果を発揮したと考えられる[4]．これらは盛り上がり箇所の誤推定を減少させるため，適合率の向上に効果を持つと考えられる．

[4] これらの素性 (A, E, F) を用いた場合，非盛り上がり発話推定の観点からの F 値は 0.76 であり，この結果は他の素性の組み合わせより高水準であった．

最も高い再現率と F 値が得られたのは，全素性から A を除いたものであった．これには，今回の実験データにおいて，非盛り上がり発話の数 (1099 件) が，盛り上がり発話の数 (617 件) に比べて著しく多かったことが影響していると考えられる．A にはどちらのラベルにも出現しやすいような，必ずしも判定に貢献しない語彙に関する情報が多く存在する．今回のようにデータ数に差がある場合，これらの情報がたまたま非盛り上がり発話に判定するために有効な情報となる可能性がある．このため，A を用いたことによって盛り上がり発話の再現性が低下し，F 値に悪影響を及ぼしたと考えられる．これを踏まえて，語彙の出現に関する素性を組み合わせる際には，語彙の意味や役割を吟味し，適切に選択して用いる必要があると考えられる．

4 おわりに

本研究では，言語情報と非言語情報の様々な情報を組み合わせて，複数人談話の盛り上がり判定を行った．その結果，盛り上がり判定の精度は，どちらか一方の情報のみに基づく精度よりも良い精度が得られた．今後の課題としては，発話行為タグや感情タグなどのパラ言語情報や，人の動作や表情などの非言語情報などの，盛り上がり判定に有効な情報を追加，選抜していくことが挙げられる．

参考文献

[1] 稲葉 通将，鳥海 不二夫，石井 健一郎，“語の共起情報を用いた対話における盛り上がりの自動判定”，電子情報通信学会論文誌．D，Vol. 94，No. 1，pp. 59-67，2011.

[2] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, “Detecting group interest-level in meetings”, Proc. IEEE International Conference on Acostics, Speech, and Signal Processing, (ICASSP’05), vol. 1, 2005.

[3] 豊田 薫，宮越 喜浩，山西 良典，加藤昇平，“発話状態時間長に注目した対話雰囲気推定”, 人工知能学会論文誌, Vol. 27，No. 2，pp.16-21, 2012.

[4] 岩波日本語表現辞典，CD-ROM，岩波書店，2002.

[5] 向井 友宏，黒澤 義明，目良 和也，竹澤 寿幸，“マイクロブログの分析に基づくユーザの嗜好とタイミングを考慮した情報推薦手法の提案”，言語処理学会第 17 回年次大会発表論文集，pp. 452-455，2011.