The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

対話におけるマルチモーダル情報を用いたユーザの興味の有無の推定

Predicting User's Interest Level in Dialogues with Multimodal Features

西本 遥人 Haruto Nishimoto　駒谷 和範 Kazunori Komatani

大阪大学産業科学研究所
The Institute of Scientific and Industrial Research (ISIR), Osaka University

We are developing a system that predicts a user's interest level for enabling natural human-system dialogues. The prediction results can be used for adaptive response generation according to the user's interest in a current topic. We incorporate multimodal features consisting of dialogue information such as switching pauses, prosody and verbal contents in user utterances, and user face images. We also develop two methods to integrate these multimodal features. The experimental results showed that using multimodal features is promising for the prediction.

1. はじめに

対話における人の発話には，その言語内容だけでなく，様々な有用な情報が含まれる．実際，対話分析においては，相手の声の韻律情報，顔の表情等の非言語情報も重要である [1]．

本研究では，人（ユーザ）とシステムの対話において，ユーザの心的状態を推定し，それを用いて応答を生成する音声対話システムの構築を目指す．本研究では，ユーザの心的状態の推定の一例として，ユーザの対話内容への興味の有無の推定を行う．

システムがユーザの心的状態に応じて応答を変えると，自然な流れで対話ができるようになる．ユーザが興味を持たない場合，システムは対話を別の話題に切り替えたり，「興味のある話題はなんですか？」等とユーザに対して話題の決定を促す応答ができる．一方興味を持つ場合，対話中の話題を掘り下げたり，その話題に似た別の関連話題を提示できる．

複数の情報を用いてユーザの状態を推定する研究は多く存在するが [2][3]，本研究では，対話内容への興味の有無を推定する．ユーザの応答から得る情報として，声の韻律情報，発話内容，顔画像情報を用いる．それら 3 つの情報に加え，本研究では，対話情報に注目する．対話情報とは，システムとユーザの対話において，「対話」という状況から得られる情報である．そして，それら 4 つの推定結果を用いて興味の有無の推定を行うシステムの構築を目指す．

興味の有無の推定を行う際にユーザの複数の情報を用いると，ある一種類の情報のみを用いる興味の有無の推定が困難な対話であっても，他の情報により統合推定結果を修正でき，より信頼できる結果を得られる可能性がある．図 1 のように，一種類のユーザの情報のみ用いる場合と複数の情報を用いる場合を比較する．システムのある発話に対してユーザが退屈な顔をして低い声で「すごいですね」と発話し，対話に興味がない場合を考える．興味の有無の推定に発話内容のみを用いた場合，「すごいですね」という発話は肯定的な文であるため，ユーザの興味の有無の推定はしづらい可能性が高い．一方，複数の情報として「ユーザの発話内容」,「顔の表情」,「声の韻律情報」を用いる場合，退屈な顔であり低い声という情報が追加される．そこから，システムは情報を統合して興味がないと推定できる可能性が高くなる．

図 1: ユーザの興味の有無の推定に複数の情報を用いた場合の利点

2. 対象データ

2.1 使用した対話コーパス

実験にはシステムとユーザの対話が収録されたマルチモーダル対話コーパスを用いた．そのコーパスは京都工芸繊維大学で収集されたものである [4]．用いたコーパスのやりとりは，以下の例のように，スポーツや音楽等の話題に沿ってシステムが質問を行ったり，話題に関する詳細な知識を紹介したりするものである．コーパスはユーザ 1 名につき 10-15 分，1 名につき約 80 のやりとりがあり，ユーザ 5 名分のデータを使用した．やりとりに関しては次の節で詳細を述べる．

やりとり例 1
S：どんなスポーツが好きですか？
U：サッカーです.
やりとり例 2
S：2020 年に東京でオリンピックが開催されます.
U：今から楽しみです.

このマルチモーダル対話コーパスにはアノテータにより興味の有無を判断したラベルが付けられている．合計 6 名のアノテータは，収集されたデータをもとに，ユーザが対話内容に興味を持っているかどうかを，1 やりとりごとに判断する．そして「興味あり (o)」,「興味なし (x)」,「不明 (t)」,「設定トラブル等で本来の対話と無関係なやりとり (e)」の 4 つのラベルか

連絡先: 西本 遥人，大阪大学産業科学研究所 知識科学研究分野，〒 567-0047 大阪府茨木市美穂ヶ丘 8-1，:06- 6879-8416，nishimoto@ei.sanken.osaka-u.ac.jp

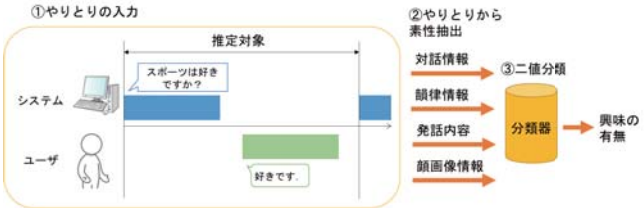

図 2: 興味の有無の推定手法の全体像

ら最適だと判断したラベル 1 つをやりとり毎に付与する．なお,「不明 (t)」のラベルは「興味あり (o)」と「興味なし (x)」の中間の意味合いをもつ．

このラベルをもとに興味の有無の正解となるラベルを多数決方式により作成した．o の数，もしくは x の数が 4 以上の場合，それを正解ラベルとした．それに該当しないやりとりのうち，e がひとつでも存在すれば e とした．これら 2 項目以外は t とした．

本実験では，興味の有無の二値分類を行うため，正解ラベルが t または e のやりとりは今回使用データに含めない．用いた正解ラベルが o と x のやりとりはそれぞれ 131，170 であり，合計が 301 であった．すなわちマジョリティベースラインは（全て x と推定した場合）は，$170/301 = 0.565$ である．

2.2 興味の有無の推定の対象区間

本研究では，システムの発話開始時刻から次のシステムの発話開始時刻までの区間を推定対象とする．図 2 の左のように，システムとユーザの発話が交互に行われる対話を想定するため，その推定対象区間にはユーザの発話が含まれている．以降，この区間を以降「やりとり」と呼び，興味の有無の推定対象とする．

3. マルチモーダル情報を用いた統合推定

与えられた正解ラベルをもとに機械学習で分類器を構築し，興味の有無の推定を行う．構築には，複数のやりとりを学習データとして用いる．図 2 に機械学習を用いた興味の有無の推定手法の全体像を示す．入力として，ユーザのマルチモーダル情報についての素性を抽出する．学習データにより構築された分類器を用い，ユーザのマルチモーダル情報を統合した推定結果を出す．推定結果は興味の有無の二値である．

以下では，対話情報，声の韻律情報，発話内容，顔画像情報の 4 つの情報で設計した素性を述べる．

3.1 対話情報の素性

対話という状況から得られる情報として以下の 4 つを用いる．ユーザの声の情報や顔の情報だけではなく，対話情報を用いて推定を行うことで，推定性能の向上を図る．

- 応答時間
- システムの発話内容の単語数
- ユーザの発話内容の単語数
- システムとユーザの発話内容の単語数の差

一つ目の応答時間は，1 やりとり中のシステムの発話終了からユーザの発話開始までの時間である．これは興味の度合いに応じて応答時間に変化が生じると予想される．これはユーザが興味を示さなかった場合より，興味を示した場合の方がユーザの応答時間が短い傾向にあるからである [5]．二つ目から四つ目の素性には，当該やりとり中のシステム発話とユーザ発話の単語数を使用する．ユーザが興味のある話題は，単語数は多くなりやすく，一方で興味のない話題は，その話題についての知識をあまり保持しないため，発話は短くなりやすいと考えられる．

表 1: 韻律情報の素性項目

素性項目の名称		素性項目の説明
RMSenergy		音量の二乗平均平方根値
mfcc	[1]	1 次メル周波数ケプストラム係数
	[2]	2 次メル周波数ケプストラム係数
	…	…
	[5]	5 次メル周波数ケプストラム係数
F0		基本周波数
voiceProb		その時点での音が声である確率
zcr		波形のゼロ交差率

表 2: 韻律情報の演算項目

演算の名称	演算の説明
max	データ中の最大値
min	データ中の最小値
range	最大値と最小値の差分
maxPos	最大値を出力した位置
minPos	最小値を出力した位置
amean	算術平均
linregc1	線形近似の勾配度
linregc2	線形近似のオフセット
linregerrQ	線形近似の二乗誤差
stddev	標準偏差

3.2 韻律情報の素性

興味の有無の推定に必要な要素がユーザの声の韻律やその変化に含まれる．例えば，対話中で他の発話よりも声が大きかったり，声が高かった場合，そのやりとりにおいてユーザが対話内容に興味を持っている可能性が高い．一方，声が小さい場合はユーザが対話内容に興味を持っている可能性は低いと考えられる．

INTERSPEECH 2009 Emotion Challenge で用いられた音響的素性セット（以下，IS09）を使用すると，韻律に関する素性が取得できる [6]．IS09 で抽出される素性は，メル周波数ケプストラム係数 (mfcc)，声のパワー (RMSenergy)，ゼロ交差率 (zcr)，声である確率に関する素性 (voiceProb)，基本周波数 (F0) といった素性項目からなる．それらは複数の演算項目によって演算される．

本推定で使用する韻律に関する素性数は 90 である．演算項目は 10 項目，素性項目は 9 項目を用いる．表 1 と表 2 は，それぞれ用いる素性項目，演算項目の一覧である．

興味の有無の推定の対象区間であるやりとり中のユーザの発話部分の音声ファイルから IS09 を用いて韻律情報に関する素性を抽出する．入力はやりとりの音声からシステムの発話区間音声を取り除いた音声を用いた．使用するコーパスにはユーザ発話のないやりとりが存在するため，その場合は素性値を全て 0 とした．

3.3 発話内容の素性

発話内容をユーザの興味の有無の推定にはユーザの発話内容をテキストに変換したものを入力とし，単語の原形に関する bag-of-words で取得する素性と，品詞の数に注目して設計

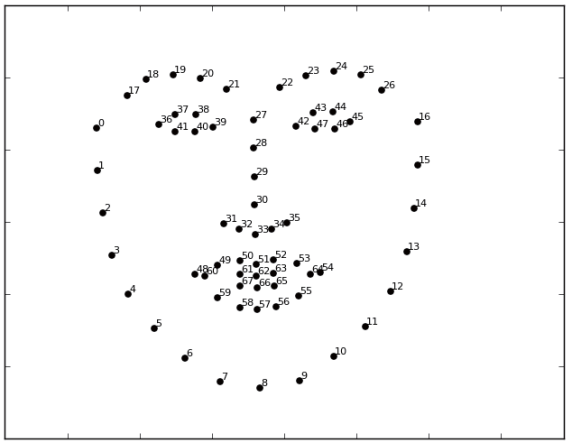

図 3: 人の顔におけるランドマークの位置

した 4 つの素性を抽出する．bag-of-words で用いる単語は登場する全ての単語から付属語とコーパス中で 1 回しか登場しない単語を除いたものである．品詞の数に注目して設計した 4 つの素性は，名詞の数，副詞の数，形容詞の数，感動詞の数である．これらは興味のある場合，発話が長くなり数が増加する傾向があると考えたため，素性に含めた．なお，形態素解析には MeCab[*1] を用いた．発話内容に関する素性数は合計 99 である．

3.4 顔画像情報の素性

顔の各部位の動き方がユーザの興味の有無に関係していると仮定し，人間の顔に含まれる特徴点（以下，ランドマーク）の位置の変化を素性として抽出する．これまでの研究で，顔の表情の分析には顔の各部位の動きが関係していることは明らかになっている [7]．ユーザの興味の有無は顔の表情に現れると考えられるので，ランドマークを用いた素性を抽出する．今回は対象となるデータが動画データであるため，その動画のフレーム毎 (100ms) にランドマークを取得する．ランドマークの位置は，静止画中のピクセル座標 (x 座標，y 座標) である．また位置の変化はピクセル座標のユークリッド距離を用いて表現する．ランドマークは Dlib[*2] を用いて取得した．

人間の顔の 68 のランドマークのうち，顔の特定の部位のランドマークのみを用いて素性抽出を行った．顔の部位の中でも口，顎は発話時や顔の表情を変化させる際に他の部位に比べてより位置が変化すると考えた．用いるランドマークは口の 4 点，顎の 1 点であり，ランドマークの動きは 12 の演算により処理した．演算は韻律情報の素性抽出に用いた表 2 の 10 の演算項目に加え，尖度と歪度を含めた 12 の演算項目である．それらは移動距離の最大値，変化の割合等である．図 3 のランドマークのうち，口には 48，51，54，57 の点を用い，顎には 8 の点を用いる．合計の素性数は $(4+1)*12=60$ である．

3.5 統合推定手法

ユーザのマルチモーダル情報をそれぞれ用いて抽出した素性を用いて，それらの情報を統合する．その手法として，late fusion と early fusion を設計した．late fusion と early fusion の統合の概略図を図 4 に示す．

*1 http://taku910.github.io/mecab/
*2 http://dlib.net/

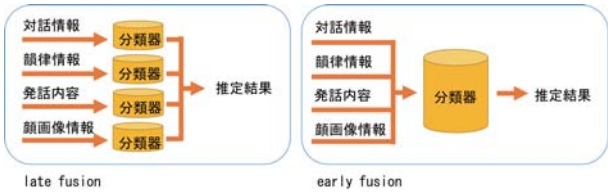

図 4: late fusion と early fusion の推定手法の概略

late fusion は図 4 のように，ユーザのマルチモーダル情報の素性をそれぞれ用いて分類器を複数作成し，その各結果を用いて統合する．統合の際には，マルチモーダル情報を用いたそれぞれの推定での SVM の判別平面からの距離に符号を与えたものを素性として，再度 SVM で分類する．

$$d = \boldsymbol{w}\boldsymbol{x} + b \tag{1}$$

式 1 の d は SVM の判別平面からの距離に符号を与えたものである．$\boldsymbol{w}$ は素性の重みベクトルであり長さは素性数に一致する．$\boldsymbol{x}$ は素性ベクトルである．b はバイアス項である．素性の重みベクトルとバイアス項は SVM による学習により得られる．与える符号は，分類器による推定結果が o の場合は正，x の場合は負とする．

early fusion は図 4 のように，各情報から得られた素性をまとめて，分類器を一つ作成し推定を行う手法である．対話情報の素性は 4，韻律情報の素性は 90，発話内容の素性は 99，顔画像情報の素性は 60 であった．

4. 実験結果と考察

表 3: 一種類の情報のみ用いた推定の正解率

発話	韻律	顔画像	対話	正解率
○	-	-	-	0.748
-	○	-	-	0.661
-	-	○	-	0.585
-	-	-	○	0.668

表 4: 統合推定の正解率

発話	韻律	顔画像	対話	early	late	oracle
○	○	-	-	0.731	0.754	0.857
○	-	○	-	0.748	0.678	0.904
○	-	-	○	0.738	0.654	0.824
-	○	○	-	0.668	0.761	0.884
-	○	-	○	0.698	0.754	0.834
-	-	○	○	0.688	0.694	0.854
○	○	○	-	0.728	0.751	0.927
○	○	-	○	0.728	0.751	0.870
○	-	○	○	0.728	0.688	0.920
-	○	○	○	0.701	0.744	0.910
○	○	○	○	0.734	0.744	0.934

ここでは，一種類の情報の用いた推定結果と複数の情報を統合した結果を比較し，情報を統合することの有用性を説明する．一種類の情報を用いた推定は，3.5 節で述べた統合手法は用いることはできないため，その一種類の情報の素性をそのまま用いて分類器を構築し，推定結果を出す．

4.1 一種類の情報のみ用いた推定結果

表 3 は一種類の情報のみを用いたユーザの興味の有無の推定の正解率である．発話内容のみを用いた推定の正解率は 0.748 であり，4 つの推定のうち最も性能がいいことがわかる．以降は，対話情報，韻律情報，顔画像情報を用いた推定の順に性能が良いことがわかる．今後は性能の低い顔画像情報を用いた推定の性能の向上が課題である．

4.2 自動的に統合した場合

表 4 は，複数の情報を用いた統合推定の正解率を示している．○のついたユーザ情報を用いて late fusion，early fusion，次節で述べる oracle で統合推定を行ったときの正解率を示している．表 3 より，発話を用いた推定は正解率が 0.748 であったのに対し，表 4 より，late fusion に注目すると，発話+韻律を用いた推定は正解率が 0.754 であった．このことから，late fusion で統合することで性能がわずかに上昇することがわかる．情報を統合することで性能は良くなるが，この結果は 2 つの情報を統合させたものである．3 つもしくは 4 つの情報を統合させた結果は，発話+韻律+顔画像，発話+韻律+対話を統合させた推定の正解率 0.754 が最高であり，2 つの情報を統合させた結果より性能が低い．そのため，より多くの情報を統合する推定の性能の向上が今後の課題である．

late fusion による 2 つの情報の統合について，組み合わせる情報によって正解率の増減に大きな差がある．例をあげると，韻律のみの正解率は 0.661，顔画像のみの正解率は 0.585 であったが，これらを統合すると，正解率は 0.761 まで増加した．同じく韻律と対話をそれぞれ用いた場合の正解率は 0.661，0.668 であったが，統合すると正解率は 0.754 に増加した．反対の例をあげると，発話と対話をそれぞれ用いた場合の正解率は 0.748，0.668 であったが，統合すると正解率は 0.654 に減少した．すなわち，正解率が増加する組み合わせは互いにあまり相関のない情報を持ち，一方が他方を補うことができるといえる．反対に，正解率が減少する組み合わせは互いに相関のある情報を持つといえる．統合することで性能を上昇させるために，互いに相関の少ない素性を設計する必要がある．

early fusion による統合では，情報を複数用いた場合でも性能の向上は認められなかった．2 つ，3 つ，4 つの情報を用いた統合の最も正解率の高い数値は，それぞれ 0.748，0.728，0.734 であり，発話のみを用いた推定の正解率 0.748 を上回らなかった．唯一，用いる情報を 3 つから 4 つに増やした場合では，性能が向上した．各情報の素性数の次元に差がある状態で統合したため，今後は素性数の差を減らすための手法を考える必要がある．

4.3 統合が理想的に行われた場合

推定性能の上限を知るために，ユーザの興味の有無の推定において，それらを理想的に統合した場合 (oracle) を考える．oracle では，正解が o である場合に，いずれかの推定結果に o が含まれれば o とする．それ以外の場合は x とする．

oracle での統合から正解率を出し比較する．oracle の正解率は，統合の性能の上限の値となる．統合推定の結果を示した 4 に注目すると，2 つ，3 つ，4 つの情報を用いた oracle による統合の最も正解率の高い数値は，それぞれ 0.904，0.927，0.934 であった．これから，より多くの情報を用いた統合推定の方が推定性能の上限が高くなるということがわかる．

5. おわりに

本研究では，ユーザから得られるマルチモーダル情報から素性を抽出した．各情報に関する素性からユーザの対話内容への興味の有無を推定した．マルチモーダル情報を統合し，ユーザの対話内容への興味の有無の推定手法を示した．実験の結果から，提案した late fusion を用いて自動的に統合した際，一種類の情報のみを用いた推定より性能が向上する場合が認められた．対話情報を含めたマルチモーダル情報を組み合わせることで，推定性能の上限が向上することがわかった．

今後の研究課題として，推定性能の向上のための素性抽出手法の改良が挙げられる．

参考文献

[1] Mark L. Knapp, Judith A. Hall, and Terrence G. Horgan. Nonverbal communication in human interaction. *Cengage Learning*, 2013.

[2] 岡田将吾, 松儀良広, 中野有紀子, 林佑樹, 黄宏軒, 高瀬裕, 新田克己. 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.

[3] 上村譲史, 目良和也, 黒澤義明, 竹澤寿幸. 字句情報，音響情報，表情から推定した話者感情の食い違い状況の分析と食い違い自動検出手法の提案. 第 78 回全国大会講演論文集, 第 2016 巻, pp. 321–322, mar 2016.

[4] 荒木雅弘, 冨増紗也華, 中野幹生, 駒谷和範, 岡田将吾, 藤江真也, 杉山弘晃. マルチモーダル対話データの収集と興味判定アノテーションの分析. *SIG-SLUD*, Vol. B5, No. 02, pp. 20–25, oct 2017.

[5] 河原達也, 川嶋宏彰, 平山高嗣, 松山隆司. 対話を通じてユーザの意図・興味を探り情報検索・提示する情報コンシェルジェ. 情報処理, Vol. 49, No. 8, pp. 912–918, 2008.

[6] Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[7] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, 1978.