2O4-OS-23a-1

雑談対話におけるマルチモーダル情報からの興味の有無の判定

Multimodal Classification of Users' Interests on Topics in Chat-oriented Dialog

冨増 紗也華 *1　　荒木 雅弘 *2
Sayaka TOMIMASU　　Masahiro ARAKI

*1京都工芸繊維大学 情報工学課程
Department of Information Science , Kyoto Institute of Technology

*2京都工芸繊維大学 情報工学・人間科学系
Information and Human Sciences , Kyoto Institute of Technology

We can enjoy long-term conversations with a robot if it uses the information of users' interest. In this paper we propose a method of classifying whether the user has an interest on topics based on multimodal information such as facial expressions and prosodic information of the user's utterance. The advantage of using this information is that we can judge the presence of interest on topics without linguistic information. In the experiment, we selected both visual information and prosodic information as a feature. Visual information consists of users' facial expressions and gestures. Prosodic information can be gotten from the user's voice. The result showed that using visual information, accuracy was 62% and using prosodic information, accuracy was 68%. Moreover, when we use both kinds of information, accuracy became 70%. We found the possibility that the performance can be improved by combining the both kinds of information.

1. はじめに

近年，さまざまな話題に関して，ユーザと対話を行うことができる雑談対話システムの研究・開発が進んでいる．システムがユーザの興味の有る話題や興味の無い話題を理解してくれており，それに合わせて話を振ってくれればユーザにとってより長期的に使いたいと感じられるシステムになるのではないだろうか．しかし，興味の有無の判定をシステムが行う場合，言語情報のみでの判定は難しいという問題がある．

例として図 1 のような状況を挙げる．この例ではシステムの「新しいドラマが今週末から始まるようですね？」という問いかけに対して，ユーザは「そうですね」としか答えておらず，文字だけでは興味の有無の判定が難しい．しかし，ユーザの表情を見ると笑顔であることから興味があると判定出来る．このように言語情報では興味の有無を判定することが難しい場合には，表情のような言語情報とは独立したマルチモーダル情報を用いてユーザの状態を把握し，興味の有無の判定を行わなければならない．しかし現在，興味の有無がどのマルチモーダル情報を用いることで判定できるのか明らかでないため，どのマルチモーダル情報に着目するかについて検討する必要がある．

どのマルチモーダル情報によって興味の有無を判定できるかが特定できれば，言語情報を用いて興味の有無を判定する必要がなくなるため，話の内容に依存することなく興味の有無を判定できる．

2. 興味の有無の判定

マルチモーダル情報を用いた興味の有無の判定をするにあたって問題となるのが，どのぐらい細かい単位で興味の有無を判定をするのが良いかという点である．

図 2 のような具体例を挙げる．話題の切れ目はひとつの候補であるが，「ドラマ」という話題には興味があるが，特定の「出演俳優」には興味がない場合など，話題の粒度設定の問題もあり，単位としては大きすぎると思われる．

連絡先: 冨増 紗也華，京都工芸繊維大学情報工学課程，〒606-8585 京都市左京区松ヶ崎橋上町 1，075-724-7125，tomi@ii.is.kit.ac.jp

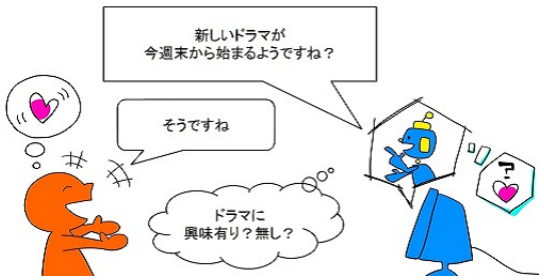

図 1: 言語情報のみでは興味の有無の判定が難しい例

そこで，興味の有無を判定する単位を事前に「ドラマ」のような話題として設定して興味の判定をするのではなく，システムの発話に対するユーザの応答という 1 交換毎に興味の有無の判定をする必要があるのではないだろうか．興味の有無を判定する単位を 1 交換毎にすることによって話題に依存することなく，よりユーザにとって興味のあるものを細かく判定することができる．

ここで興味の有無を判定する単位である交換について述べる．一つの話題を

$$d = t_1, t_2, \ldots, t_i, \ldots, t_n \tag{1}$$

と表す．ここでの d は一つの話題を表し，n 個の交換から構成される． t_i は話題中の i 番目の 1 交換を表す．ここでの対話の 1 交換は，システムが行う発話に対するユーザの応答との組み合わせで成るものとする．交換 t_i におけるシステムの発話ターンを S_i，それに対するユーザの応答ターンを U_i とし，

$$t_i = S_i + U_i \tag{2}$$

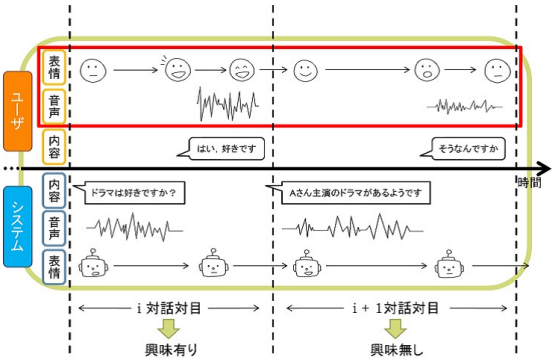

図 2: システムとの対話の具体例とマルチモーダル情報の変化

と表す．この 1 交換内におけるユーザの状態を特徴としてとらえ，興味の有無を 1 交換毎に判定する．この時特徴を

$$\boldsymbol{f_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}, \ldots, x_{im}) \tag{3}$$

と定義する．式における $\boldsymbol{f_i}$ は交換 t_i における特徴全体を表し，x_{ip} は交換 t_i における p 番目の特徴を表す．対話中における交換は時系列になっているため，それぞれの交換における特徴も系列的に表現でき，特徴の系列を F とすると

$$F = \boldsymbol{f_1}, \boldsymbol{f_2}, \ldots, \boldsymbol{f_i}, \ldots, \boldsymbol{f_n} \tag{4}$$

と表せる．これらの特徴に対する興味の有無の判定結果を

$$R = \boldsymbol{r_1}, \boldsymbol{r_2}, \ldots, \boldsymbol{r_i}, \ldots, \boldsymbol{r_n} \tag{5}$$

として定義する．ここでの $\boldsymbol{r}$ は $\boldsymbol{f}$ のそれぞれの要素に対応した興味の有無の判定結果である．このように定義することによって系列 $\boldsymbol{f}$ に対するラベル列 $\boldsymbol{r}$ の系列ラベリング問題として考えられる．しかし，今回の場合は興味の有無の判定であるため，ある交換における興味の有無はそれ以前の交換における興味の有無に影響を受けにくく，出力系列の依存性が低いと仮定する．判定結果系列と対話内容からの具体的な話題に対する興味の有無の判定は今後の課題とし，今回は 1 交換毎で興味の有無の判定を行う．これらをまとめた図として図 3 を示す．

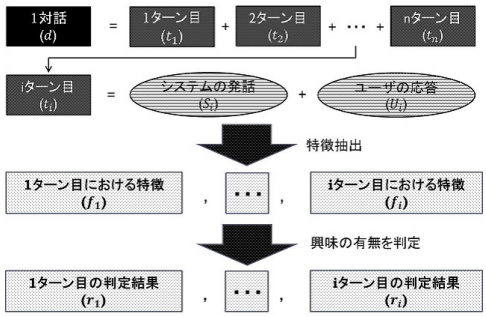

図 3: 興味の有無の判定手順

3. 実験

使用するマルチモーダル情報として，表情や身振りといった情報を視覚的特徴，ユーザの音声の韻律情報を聴覚的特徴と定義して識別を行った．またこの 2 つを用いた特徴における識別の比較と識別精度についての考察を行った．

3.1 実験方法

システム側として MMDAgent[*1] を用い，実験参加者は大学生 10 名 (男性 5 名，女性 5 名) で実験を行った．実験における流れは，まず実験参加者に事前アンケートに答えてもらい，あらかじめ用意した 15 種類の話題から興味の有る話題と興味の無い話題をそれぞれ 3 種類ずつ選んでもらった．数日後 MMDAgent との対話を行い，実験終了後，対話中に出現した 6 つの話題に対する詳細なアンケートに答えてもらった．

実験における話題の展開方法としては，一つの話題につき 9 〜15 交換を一区切りとし，興味の有無に関わらず話題の展開方法を同じようにするために「その話題に対しておすすめのものはあるのか」を尋ね，その後に「その話題に関する最新情報などの情報の提供」をするものに固定して実験を行った．実験参加者ごとに事前アンケートから選ばれた 6 つの話題の順序はランダムに並び替えて行った．実験終了後のアンケートとして，ユーザのそれぞれの話題に対する詳細な興味の度合いを知るため，実験において取り上げた話題に対しての興味の度合いと実際に実験において話された内容についての興味の度合いについてそれぞれ話した話題について 4 段階の評価をしてもらった．

実験参加者の様子を記録するために，ディスプレイの上からビデオカメラによる上半身の撮影と Microsoft 社製 Kinect for Windows[*2] による上半身の撮影を行った．また，音声に関しては実験参加者の近くに置かれたマイクからの音声と MMDAgent の音声を収録した．

実験中，実験参加者には椅子に座ってもらい，ディスプレイと実験参加者の間は 80cm 程離して実験を行った．MMDAgent は画面上に表示され，実験参加者からはノートパソコンに繋がっているマイクによってエージェントとの対話を行っているように見えるようにした．ノートパソコンはインターネット通話でシステムのふりをした人間 (Wizard) 側である実験者のパソコンと音声通話状態であり，ノートパソコンに繋がっているマイクによって実験者に音声が届くようになっており，その音声を聞いて臨機応変にシステム側の発言を行った．

(1)WOZ 法を用いたエージェントの発話システム

システム側からの実験参加者への発話は，WOZ 法を用いるにあたって自然な対話をするためにユーザの発話に合わせて臨機応変に出来るだけ言語的破綻のないように発話させなければならない．そのため，WOZ 法を用いたエージェントの発話システムを作成し，MMDAgent に予め用意しておいた発話を選択し，送信することによって発話を行えるようにした．実験において実験者が使用するエージェントの発話システムとして，図 4 のようなシステムを作成した．

このエージェントの発話システムは MMDAgent の発話内容を統一することと，ユーザからの発話とシステムの発話との間を開けすぎないことを目的として作成した．このシステムは予め文とその文のジャンルを設定しておくことでジャンルを選ぶとそのジャンルに関する文集合を表示し，選択することによって MMDAgent の発話内容を選択するシステムである．

また，拡張機能として文の末尾に特定の文字列を入力することによって「喜び」「疑問」「紹介」「微笑み」の 4 パターンの表情や動作で MMDAgent が変化するようになっている．

(2) 着目したマルチモーダル情報

マルチモーダル情報において着目する点は，視覚的特徴として顔表情と頭部の動作の変化，聴覚的特徴としてユー

*1 http://mmdagent.jp/
*2 https://dev.windows.com/en-us/kinect

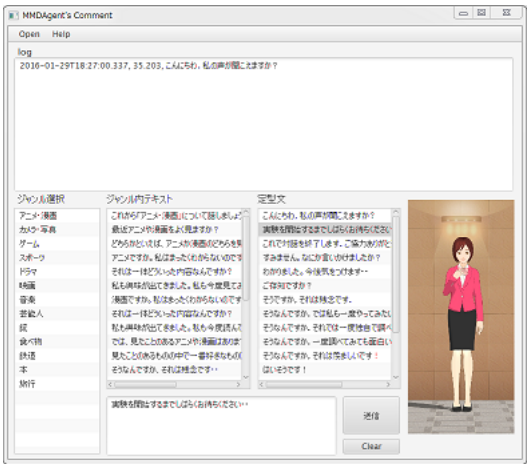

図 4: WOZ 法を用いたエージェントの発話システムのインタフェース (Copyright 2009-2016 Nagoya Institute of Technology (MMDAgent Model “ Mei ”))

ザの応答時の音声の変化に着目した．視覚的特徴は Microsoft 社製 Kinect SDK，聴覚的特徴はオープンソフトウェア openSMILE[Eyben 2010] を用いて特徴量の取得を行った．

Kinect SDK によって分析される特徴の中から顔表情として笑顔 (笑顔であるかどうか)，頭部の動作の変化として回旋角度 *3，屈曲・伸展角度 *4，側屈角度 *5 に着目して特徴の抽出を行った．これらの特徴を，システムが発話しているターンにおける特徴とユーザが発話している特徴に分けてそれぞれのターンにおいて 15 次元の特徴の出力を行った．よって 1 交換における特徴はそれぞれのターンにおける 15 次元の特徴を合わせた 30 次元として興味の有無の識別を行った．それぞれのターンにおいて具体的に抽出した特徴を表 1 に示す．表 1 における Unknown は顔の角度などによって認識が上手くされず，値が取れなかった時に出力される値を指す．

表 1: 視覚的特徴としてターン毎に抽出した特徴

取得可能な特徴	出力値
笑顔	笑顔であると判定されたものの割合 笑顔でないと判定されたものの割合 Unknown と判定されたものの割合
回旋角度	最大値 最小値 変位 (最大値と最小値の差) 平均値
屈曲・伸展角度	最大値 最小値 変位 (最大値と最小値の差) 平均値
側屈角度	最大値 最小値 変位 (最大値と最小値の差) 平均値

openSMILE はユーザの音声による韻律特徴から得られる基本周波数のような特徴量とそれぞれの特徴量における最大値などの素性値が取得可能である．これらの組み合わせによって得られた 384 次元の特徴を使って興味の有無の識別を行った．

このようにして得られた特徴から，視覚的特徴として 30 次元の特徴による識別と，聴覚的特徴として 384 次元の特徴による識別，またこれら 2 つの特徴を組み合わせた合計 414 次元の特徴による識別を行った．

3.2 評価手法

本研究においては SVM を用いて識別を行い，leave-one-out 法による評価を行った．今回の場合は個人によって特徴が変化すると考えられるため実験参加者毎に識別を行った．

興味の有無のアノテーションは大学生 3 名 (男性 1 名，女性 2 名) が収録した動画を見て 1 交換毎のアノテーションを行った．アノテーションは視覚的，聴覚的，言語的に見て各実験参加者がその交換における話や質問に対して興味が有るのかどうかという観点で Yes(実験参加者がその交換において興味を示していると判定できる)，No(実験参加者がその交換において興味を示していないと判定できる)，Neutral(実験参加者がその交換において興味の有無の判定が難しいく，判定できない) のアノテーションをつけてもらった．このようにしてそれぞれの交換に対してアノテーションを付けてもらい，3 名のアノテーション結果から多数決を取ることによって最終的なアノテーションを決めた．3 人のアノテーション結果が 3 つに割れた場合はその交換は判定が難しい交換であったとし，Neutral の判定とした．今回の実験においては Neutral は興味が無いとして No と同じカテゴリとし，Yes と No の 2 値による識別を行った．アノテーション結果の合計数は Yes が 48%，No は 52%となった．

また，実験におけるシステムと実験参加者が正しく対話ができていないと判定できる交換 (実験参加者がシステムの発話を聞き取れず，聞き直した交換など) は今回の実験においては想定外の交換とし，識別の際には用いず識別を行った．

3.3 結果

実験参加者毎に識別を行って得られた識別結果の平均値を表 2 に示す．結果より，視覚的特徴を用いた場合の識別率は 62%，聴覚的特徴を用いた場合の識別率は 68%，視覚的特徴と聴覚的特徴の両方を用いた場合の識別率は 70%であったことがわかる．よって視覚的特徴と聴覚的特徴を組み合わせることによって興味の有無判定における識別精度が上がる可能性が考えられる．

表 2: 実験参加者の識別結果平均値

	特徴	正解率	分類	適合率	再現率	F 値
	視覚的		興味有り	0.60	0.50	0.53
			興味無し	0.61	0.71	0.65
平均		0.62		0.61	0.62	0.61
	聴覚的		興味有り	0.67	0.65	0.66
			興味無し	0.68	0.70	0.69
平均		0.68		0.68	0.68	0.68
	両方		興味有り	0.69	0.67	0.68
			興味無し	0.70	0.72	0.71
平均		0.70		0.70	0.70	0.70
※平均は重み付き平均						

*3 首の骨を軸とした時の軸に対する回転角度)
*4 屈曲角度 (頭を前方に傾ける方向の角度)，伸展角度 (頭を後方に傾ける方向の角度
*5 首を側方に曲げる角度

3.4 考察

アノテーション結果の妥当性について考える．実験において行った実験参加者へのアンケート結果と 3 名によるアノテーション結果から，実験において取り上げた話題に対しての興味の度合いと実際に実験において話された内容に対しての興味の度合いでそれぞれ分類とアノテーションが一致している交換数を全交換数で割った数字の平均値を求めることによって，人間がどの程度興味の有無を判定できるのかを調べた結果，今回の実験においては 10 名の実験参加者の平均値から 60%から 84%の精度であることがわかった．よって視覚的特徴における識別，聴覚的特徴における識別，また両方を用いた識別において人間と同等程度の精度が得られることがわかった．以上のことから視覚的特徴と聴覚的特徴を用いることによって人間と同等程度の識別精度が得られ，さらに 2 つの特徴を組み合わせる事によって識別精度が高くなることがわかった．

4. 関連研究

音声情報と興味判定の研究として倉野ら [倉野 2013] の研究がある． 倉野らは他の作業をしながらの動画の閲覧に着目し，閲覧時のユーザの興味の変化を推測した．映像閲覧デバイスとしてタブレット端末を用い，マルチモーダル情報として音声情報，画像情報，三軸加速度を用いて実験を行った．この研究においては興味推定の仮説を設け，それについて検証した実験を行った．映像に対して閲覧者が何か反応した際，三軸加速度と音声データに変化が生じる，複数人で映像を閲覧する際，その発話や行動によって映像に対して興味を持ったポイントを推定できる，映像を閲覧するシチュエーションによって，閲覧者の反応の仕方に変化が生じるという三点においてそれぞれ妥当な仮説であるという結論を出している．本研究との違いとしてはユーザの興味を向ける対象に違いが有る．しかし対象に違いがあったとしてもこれらの仮説の検証は本研究においても有意義なものであり，これらの変化を推定するものとしてマルチモーダル情報の中でも視覚的特徴と聴覚的特徴に着目して実験を行った．

また，韻律情報からの興味判定の研究として中村ら [中村 2015] の研究がある．中村らは擬人化エージェントとの音声対話におけるマルチモーダル情報から対話内容に対して難しいと感じたか否かと興味を持っていたか否かを推定する研究を行っている．用いる画像情報としては視線，表情，姿勢，手振りの 4 つの非言語動作から推定をしている．本研究との違いとしては画像情報の選択方法に違いがあり，画像情報としてより頭の部分における変化に着目した研究を行った．

さらに，ユーザの対話意欲を推定するためにマルチモーダル情報を用いた研究として千葉ら [千葉 2015] の研究がある．千葉らは雑談対話中のユーザの対話意欲を推定するために言語情報，音声情報，画像情報をそれぞれの組み合わせで用いた識別方法についてそれぞれ検討している．本研究との違いとしては，人-人間の対話を扱うのではなく，人-システム間での対話を扱ったこととマルチモーダル情報の抽出方法においての違いが挙げられる．本研究においては画像情報においてはより顔周辺の変化に着目した特徴，音声情報においては openSMILE を用いる事によって，openSMILE の今後の精度向上の可能性を考慮した特徴の取得を行った．

5. おわりに

今回の研究ではシステムとの雑談対話においてユーザの興味の有無の判定を行うことに対して，マルチモーダル情報として視覚的特徴と聴覚的特徴を用いた興味の有無の判定を行った．結果として視覚的特徴と聴覚的特徴の両方において興味の有無の判定することが可能であり，これら 2 つの特徴を合わせて用いる事によって判定の精度が向上する可能性があることがわかった．

今後の課題と展望として，実験参加者の中に「自分の好きなものを聞かれるのはあまり得意ではないので答えにくかったが，提示される情報に対しての意見は言いやすかった」と回答した人がいたことから，システム側からの話題の展開方法における「その話題に対しておすすめのものはあるのか」と「その話題に関する最新情報などの情報の提供」のどちらの方が興味の有無を判定しやすいのかといった点で考察を行っていきたい．また，今回の研究では表情における観点が笑顔であるかどうかでしか判定できなかったため，より細かく表情を考慮した研究を行っていきたい．

参考文献

[Eyben 2010] Eyben, F., Wöllmer, M., and Schuller, B.: openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor., *Proceedings of the 18th ACM international conference on Multimedia*, p.1459-1462 (2010).

[倉野 2013] 倉野 大二郎，松村 耕平，角 康之: マルチモーダルデータを用いた映像閲覧者の興味推定，情報処理学会，p.435-439 (2013).

[千葉 2015] 千葉 祐弥，伊藤 彰則： ユーザの対話意欲推定のための人対人対話データの分析と WOZ システムの検討，研究報告音声言語情報処理（SLP），p.1-6 (2015).

[中村 2015] 中村 和晃ら： 擬人化エージェントとの音声対話時におけるユーザの非言語動作からの難/易及び興味/退屈の推定，研究報告音声言語情報処理（SLP），95，p.85-96 (2015).