複数人ユーザとエージェントの会話における受話者推定の誤り時の状況の考察

Consideration of the situation of the Errors in Estimating the Addressee of User Utterances in Multi-user Agent Conversation

堀田 怜[1*]　乙木 翔地[1]　黄 宏軒[1]　川越 恭二[1]
Ryo HOTTA[1] Shochi OTOGI[1] Hung-Hsuan HUANG[1] Kyoji KAWAGOE[1]

[1] 立命館大学情報理工学研究科
[1] Graduate School of Information Science and Engineering, Ritsumeikan University

Abstract: Nowadays, embodied conversational agents are gradually getting deployed in real-world applications like the guides in museums or exhibitions. In these applications, it is necessary for the agent to identify the addressee of each user utterance to deliberate appropriate responses in interacting with visitor groups. However, as long as the addressee identification mechanism is not completely correct, the agent makes error in its responses. Once there is an error, the agent's hypothesis collapses and the following decision-making path may go to a totally different direction. We are working on developing the mechanism to detect the error from the users' reactions and the mechanism to recover the error. This paper presents the first step, a method to detect laughing, surprises, and confused facial expressions after the agent's wrong responses. This method is machine learning base with the data (user reactions) collected in a WOZ (Wizard of Oz) experiment and reached an accuracy over 90% .We plan to gather and analyze the situations where the addressee estimation mechanism tends to make errors from the corpus collected in previous study. Then we would develop the error detection method which also consider those situations from the analysis results.

1　はじめに

今日，擬人化会話エージェントが登場してきている．擬人化会話エージェントは，画面に人間を模したキャラクタを表示させ，そのキャラクタと会話を行えるものである．しかし現状では，それらはユーザ一人に対して会話を行うものが多い．エージェントが設置される展示会や博物館などでも，単独のユーザだけでなく家族やグループユーザにシステムが利用されることも想定される．

複数人に対してエージェントが会話を行う時に，一人との時は起こらなかった，受話者推定の必要性が発生する．受話者推定とは，誰かが発言した時に，その発話が誰に対して発せられたのかを推定することである．この機能を実装しなければ，エージェントはユーザの発話に対して全て発話を返すので，ユーザ同士が会話を行いたい時でも反応してしまい，会話が成り立たなくなってしまう．また，ユーザの意図しないほうに会話が進んでしまう可能性もある．これまでの受話者推定の先行研究 [1][2][3] で，非言語情報や言語情報を用いて受話者推定を行っており，それぞれ約 80%，60%，F 値 0.72 の精度を報告している．いずれも 100% 認識を行えておらず，人間同士の会話でも受話者推定の誤りは起きるので，この機能は 100% 成功させることはできないと考えられる．受話者推定を誤った時の対処をしなければ一度誤った時にその後の会話がずれ始める可能性がある．

観光案内の文脈で，福岡県について会話を行っており，ユーザ A がユーザ B に発話したが，エージェントが反応してしまった場合の例を以下に挙げる．

ユーザ A：ハウステンボスどうだった？
エージェント：ハウステンボスは，石畳やレンガのひとつひとつにこだわり古きよきヨーロッパを再現した街並みです．景観へのこだわりから、電柱も一切ありません．

もう一人のユーザに体験談等，その人にしか答えられない会話を行った時に，エージェントがハウステンボスについての説明を行うと，ユーザが困惑し，会話が停滞するとともに，エージェントに対して不信感を

*連絡先：立命館大学情報理工学研究科
〒 525-8577 滋賀県草津市野路東 1-1-1
E-mail: hotta@coms.ics.ritsumei.ac.jp

持ってしまう可能性がある．エージェントがユーザの質問に対してまったく反応しない場合でも同様のことが起こると考えられる．

そこで本研究は，受話者推定を誤った場合に, ユーザの表情や音声の変化などの反応からそれにエージェントが気付くシステムを実現することを目的とする．まず，エージェントが受話者推定を誤った時のユーザの反応を得るために，WOZ (Wizard of Oz) コーパス収集実験を行う．実験で得たコーパスを分析することにより，特徴となる，顔情報や音声情報を得る．エージェントが誤りに気付き，訂正することで, 会話の流れを戻すことや, エージェントが会話内容への理解能力を高めることが期待できる．

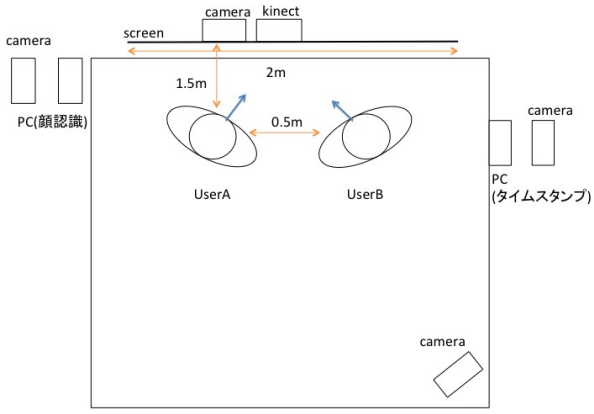

図 1: コーパス収集実験の環境

2 関連研究

複数人会話では，一対一の会話では起こらない，受話者が誰なのかの問題が発生する．人間同士では意識することなく複数人で会話を行うが，エージェントはそれを行うことを出来るようにするには，受話者を推定する機構が必要となる．Huang ら [1] の先行研究では,「顔向き」,「声の高さ」,「声の強さ」,「話速」などから受話者が誰なのかを推定する手法を提案した．また，Frampton ら [2] は音声認識や画像処理技術から受話者を推定している．さらに，Katzenmaier ら [3] は音声情報と視覚情報の統合により発話がロボットに向けられているかどうかを推定する．Bohus ら [4] は，4 つのフロアで制御を行っている．Hold Action はユーザがフロアを保持している，Release Action は他ユーザにフロアを受け渡す，Take Action はフロアを要求している，Null Action はフロアを要求していない状態で，システムにフロアが渡されたときにシステムから出力がされる．フロアの受け渡しは注視情報から行われている．

顔向きといった非言語情報は人間同士が会話するなかで用いられており，エージェントが受話者を推定する特徴量として有用な情報と判断できる．例えば，中島ら [5] は，音声対話システムに非言語情報としてユーザの顔表情の認識を用いている．Ekman[6] は喜び，悲しみ，恐怖，驚き，嫌悪，怒りの 6 種類を基本表情としている．中島らは，機械との対話ではユーザは「笑い」,「しかめ面」,「見開き」の 3 表情が多く現れるとしている．また，Ekman の悲しみや怒りは機械との対話ではほとんど表れないとしている.「笑い」は，対話ロボットがユーザの発話を誤解釈した時や，ユーザが質問をしていないのに回答を発話してしまった場合で,「しかめ面」は，会話が成立していない状況が多く続いた場合に多く現れるとしている．3 つの表情を目・眉，頬，口の 4 つの部位の情報を用いることにより判定している．

大塚 [7] は「どの行動がどの行動の引き金となっているか」といった行動間の因果関係を用いたインタラクションネットワークを呼ぶ表現法を提案している．ここでインタラクションの分析のために発話と頭部ジェスチャに注目している．頭部ジェスチャは「頷き」,「首振り」,「傾げ」を用いている．話し手が話している間，聞き手は相槌や頷きをする等，話し手と聞き手の頭部ジェスチャは重要な手掛かりであるとしている．このように様々な非言語情報が音声対話で用いられている．会話エージェントの受話者推定に使える情報は顔向き，頷き，表情等，多岐にわたるが，先行研究で 100% 推定出来ておらず，全てを使用したとしても，100% 推定できないと想定される．

そこで本研究では，このような会話エージェントが受話者推定を誤った時に，声の変化 (声の高さ，声の強さ，話速など)，向き合う状態，表情の変化といった非言語情報を用いて，会話エージェントが受話者推定を誤ったことに気付くことで会話の混乱や停滞を解決する手法を提案する．音声認識だけでは，受話者推定を誤った時に，ユーザが必ず指摘してくれるとは限らず，戸惑ったり，笑ったりするだけの可能性があり，非言語情報の表情の変化や，もう一人のユーザに確認を取るなどの姿勢の変化が起こりやすいと推測できる．また，非言語情報は個人差が比較的少なく，音声認識よりも判断しやすいと推測できる．さらに，受話者推定を誤る状況についても合わせて検討する．

3 コーパス収集

3.1 コーパス収集実験の設定

被験者は立命館大学の大学生，大学院生の友達同士の同性の二人一組で，図 1 の環境で行った．実験では，

表 1: 誤りの有無と表情変化の関係性

	無表情	笑い	戸惑い	驚き
ユーザがエージェントに話しているのにエージェントが反応しない	8 回	7 回	13 回	0 回
ユーザがもう一人のユーザに話しているのにエージェントが反応してしまう	3 回	5 回	1 回	3 回
誤りでないときの表情の変化	214 回	141 回	12 回	31 回
各表情の誤り時の確率	4.8%	7.8%	53.8%	8.8%

等身大の女性キャラクタの擬人化会話エージェントがスクリーンに投影される．4 台のビデオカメラで，正面，後ろ，顔認識，ソフトの出力画面，タイムスタンプのログを記録した．各組は擬人化会話エージェントとの会話を 3 セッション行い，3 セッション目で意図的に誤りを入力した．1 セッション目はエージェントとの会話に慣れておらず，何をすればいいか分からず戸惑うことがあると考えられるため，ユーザをエージェントに慣れさせる．2 セッション目でエージェントと慣れた上で会話を行うことで，エージェントは人間と同じように会話を行えると感じさせる．3 セッション目で誤りを入れることで，ユーザのエージェントの誤りに対する反応を得ることが出来ると期待した．タスクの偏りが無いよう，順番はすべての組み合わせで行った．実験の風景として図 2 に示す．実験実施の共通事項として，

- 二人が一緒に行いたいことについてエージェントから情報を得ながら二人で議論する
- エージェントは行えることの情報を持っており，ユーザはエージェントからそれらの情報を得ることができる
- 二人で納得した選択が決定されるまで議論を行う
- 10 分間会話を行い，第 1 希望から第 3 希望まで 3 つ選択してもらう

タスクの内容を以下に述べる．

- 履修登録 : 12 の授業の教養科目から一緒に履修するものを選ぶように教示した. 被験者は会話エージェントに質問することにより, 授業の情報を知る事ができる. 会話エージェントが持つ情報は, 授業の概要，使用する教材，単位取得の難易度，授業の時間帯などである
- 旅行計画 : 九州の観光スポット 14ヶ所のうち,3ヶ所を無料で行くことができる旅行クーポンを入手した設定で, 一緒に生きたい場所を選ぶように教示した. 会話エージェントに質問することにより, 観光スポットの情報を知る事ができる. 会話エージェントが持つ情報は, 各観光スポットについての簡単な歴史，ゆかりのある有名人，見どころ，お土産，近くにあるレストランなどである.
- アルバイト選択 : 大学周辺の店などでのアルバイト 14 種類のうち，一緒に働きたい場所を 3 つ選ぶように教示した. 会話エージェントに質問することにより, アルバイトの情報を知ることができる. エージェントがもつ情報は, 仕事内容, 時給, 時間帯, 店舗の場所などである.

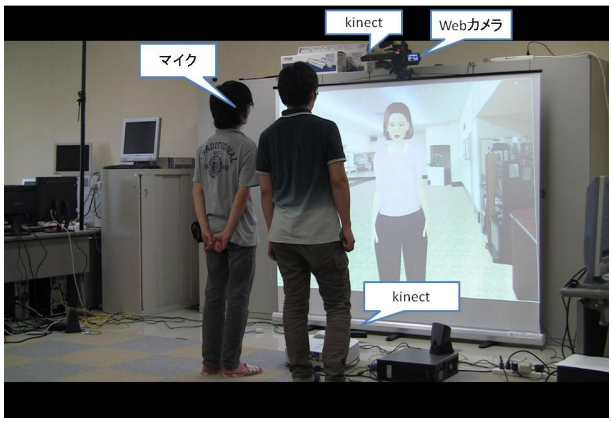

図 2: コーパス収集実験の風景

3.2 受話者推定の誤り場面の操作

受話者推定の誤りとして，以下の二つを定義した

- ユーザがもう一人のユーザに対して話しているのにエージェントが反応してしまう
- ユーザがエージェントに対して話しているのにエージェントが反応しない

実験の 3 セッション目に以下の規則で誤りを意図的に入力した.

- 約 3 分に一度の頻度で誤りの入力を行った．先行研究の受話者推定の精度が約 80% なので 5 回に 1 回程度だが，タスクや，発話内容により間隔が変わる．短い間隔で誤りをいれてしまうとすぐに誤りに慣れてしまうことにより，3 分に 1 度で誤りの入力を行った．
- 受話者推定を誤ったと気付いてもらうために，今までに答えられていた，当然答えられるような内容で行った．
- もう一人のユーザに対しての発話にエージェントが反応した場合の発話内容は, 実際の会話エージェントシステムのシミュレーションとしてエージェントが持っている発話の特徴的なキーワードをルール化し, その言葉をユーザが発話した場合に，該当する発話を行った．例えば，ユーザ A：「福岡市ってどんな食べ物があるっけ」

  と，もう一人のユーザに発話した場合,「福岡市」「食べ物」の 2 つがキーワードと想定し，それに対応する福岡市のお勧めの食べ物についての発話を行った．

4 受話者推定の誤り検知手法

本研究は，表情と音声から受話者推定の誤りを検知する．表情はユーザの目と口の開口度，眉，顔向きから，無表情，笑い，驚き，戸惑いに分類する．ユーザの平均的な声の高さ，強さ，話速を求め，普段と違う値を出した場合を調べる．

4.1 表情・音声と誤りの関連性

人間は会話内で誤りが起きたときに，戸惑ったり，驚くなどの反応を起こすと考えられる．これらの反応は表情の変化を伴うことが多い．会話内での誤りは，会話内容の誤りや，受話者推定の誤りなどがある．発話内容の誤りによる表情の変化は，ユーザが増えた複数人との対話でも起こることが考えられる．誤りに対してユーザが表情を変化させてもそれが誤りが原因で変化させたのか，会話の流れでの変化なのかは，表情の変化だけでは判別することは困難である．さらに，ユーザが複数人いるので，ユーザが何かを感じた時に，もう一人のユーザを伺うことも起きるので，顔の位置や角度の情報を検討する必要がある．

また，受話者推定を誤ると，表情の変化に加え,「え?」「ん?」といった，普段より声が高くなるなどの変化が起きる．このような音声の変化も表情とあわせることで誤り検知の精度の向上を狙う．

4.2 収集したコーパスの分析

表情が変化している間を区切り，左右それぞれのユーザの表情の変化の切り出しを行った．切り出した対話データをアノテーションツール Anvil[1] を用いてラベリングを行った．表 1 に分析の結果，誤りの有無と表情変化の関係性をまとめた．ここで，ラベルとして，平常時の表情を neutral，笑いを laughed，戸惑いを confused，驚きを surprised と定義した．誤りが起きたなかで 5 秒以内に表情の変化した数は 40 回中に 29 回だった．これにより表情の変化は受話者推定の誤りに関係があるのではないかと考えた．さらに，ラベリングの結果から以下の 3 つのことが分かった．

- 戸惑いが他の表情に対して誤りの確率が高い
- 直前にエージェントが発話を行っていないときに驚いた場合，誤りではない可能性が高い
- 直前にエージェントが発話を行ったときに戸惑った場合，誤りでない可能性が高い

ラベル付けした 8 人のラベルの数について表 2 に示す．無表情，笑いは普段の会話でも頻繁に起きるので，数が多く現れたが，驚きや戸惑いは，誤りを入力した時などによく起きていたが，普段の会話では現れることが少なかった．

表 2: 全てのラベルの数

表情	最大	最小	平均	標準偏差
neutral	35	9	21.62	7.06
laughed	24	7	15.37	4.87
confused	5	1	2.50	1.58
surprised	7	0	2.75	2.04

4.3 受話者推定の誤り検知手法

分析で得た表情と音声の情報を基に利用した特徴量について述べる．表情の特徴量として以下の設定をした．

- 平常時の表情 : ほとんどの時間がこの表情に当てはまるので，無作為に切り出した
- 笑い : 口が無表情のときよりも横に広がっている
- 戸惑い : 首をかしげる，眉をひそめる
- 驚き : 口を大きく開く，顔が後ろに下がる

[1] http://www.anvil-software.org/

音声の特徴量として発話ごとに以下の特徴量を音声分析ソフト Praat[2]を用いて解析する．

- ピッチ：音声の高低
- パワー：音声の強弱
- 話速：音声の速さ

4.4 表情の検出

Ekman と Friesen が 1978 年に発表した FACS(Facial Action Coding System)[3]は，顔の口や眉などのパーツの動きを 64 個の Action Unit (AU) として，定義をした．Action Unit を複数用いることで表情を記述できるとしている．今回の実験でも Action Unit を用いることで表情を認識できるのではないかと考えた．表情を検出するために，ウェブカメラで撮影した実験映像を，Action Unit を認識できる顔認識ソフト visage|SDK[4]を用いて解析した．

- Nose wrinkler (AU9)：鼻のしわ
- Jaw drop (AU26)：あごが下がる
- Lower lip drop (AU16)：下くちびるが下がる
- Upper lip raiser (AU10)：上くちびるが上がる
- Lip stretcher (AU20)：口の伸び
- Lip corner depressor (AU13/15)：口の端が下がる
- Outer brow raiser (AU2)：外側の眉が上がる
- Inner brows raiser (AU1)：内側の眉が上がる
- Brow lowerer (AU4)：眉が下がる
- Rotate eyes left (AU61/62)：目の動き
- Rotate eyes down (AU64)：目が下がる
- Position Distance：顔の位置の動いた距離
- Rotation Distance：顔の角度の動いた距離
- Rotation：顔の角度

これらの情報を 30fps で取得し，4.2 節で付けたラベル部分の情報を切りだした．Nose wrinkler から Rotate eyes down は，Action Unit に対応する，顔のパーツの動きの変化量を求めることができる．変化量として，例えば Jaw Drop なら，あごが下にどれだけ強く動いたかを数値で返し，あごが上がった時はマイナスで表す．Position Distance と Rotation Distance は，顔認識で得られた顔の情報を，前 10 フレームの平均と 10〜20 フレーム前の平均の差で求めた．Rotation は，実験では左右のユーザが居り，もう一人のユーザを見る時に左右のユーザで向く方向が逆になるので，左側のユーザは縦方向の x を除いた y と z 座標の値を反転した．

4.5 機械学習の結果

4.4 で求めた顔の情報を，人間が整理するには数が多すぎるので，機械学習を行う．得られた顔の情報を教師データとし，データマイニングツール Weka[5]を用いて機械学習を行った．Weka は，機械学習のアルゴリズムの集合体で，データのクラスタリングや可視化を行うことができる．機械学習により，顔の情報から今どの表情をしているかを求める．機械学習のアルゴリズムとして，複数のアルゴリズムを検証した結果一番精度が高かった，Random Forest を用いることにした．Random Forest は，教師データを複数に分け，それぞれで決定木を作成し，結果を統合し，その中の多数決でどこに分類するかを決めるアルゴリズムである．分類の結果，表 3 が示しているように，一番大きく影響を与えているのは Lip stretcher，次に Inner brows raiser だということが分かった．誤りがあるときのほうが普段よりも表情の変化が起こりやすいが，普段の表情の変化と区別がつかないので，表情だけではエラー検出を行うことが難しいと分かった．音声など，ほかの特徴量も検討している．

表 3: 各表情の検出精度

表情	適合率	再現率	F 値
neutral	0.866	0.866	0.866
laughed	0.887	0.890	0.888
confused	0.962	0.960	0.961
surprised	0.947	0.944	0.945
分類精度	90.9%		

4.6 誤りが起きやすい状況の確認

Huang らの先行研究の受話者推定のシステムを用い，受話者推定を行う.「顔向き」,「声の高さ」,「声の強さ」,「話速」を用いて受話者推定を行っている.「顔向き」は顔認識ソフト FaceAPI[6]用い,「声の高さ」,「声の強さ」,「話速」は Praat を用いている．これらの特徴量を Weka

[2]http://www.fon.hum.uva.nl/praat/
[3]http://www.cs.cmu.edu/afs/cs/project/face/www/facs.htm
[4]http://www.visagetechnologies.com/
[5]http://www.cs.waikato.ac.nz/ml/weka/
[6]http://www.seeingmachines.com/product/faceapi/

を用いることにより，受話者推定の結果を求めている．事前に正解のデータとなる発話ごとの発話者と受話者を Anvil でラベリングしておき，受話者推定の結果と照らし，誤っている箇所を見つける．先行研究の受話者推定の精度は約 80%なので，残りの誤った箇所の顔情報と音声情報を取り出し，どの特徴量の影響が大きいか共通点を探す．この共通点を見つけることで受話者推定の誤りが起きやすい状況を把握する．誤りやすい状況で，エージェントに様子見をさせるなどの慎重な行動をとらせ，さらに誤りやすい状況と，表情の変化などの誤った時の状況の両方を考慮することにより，受話者推定の誤りの検知を行いたい．

5 おわりに

本研究では，複数人ユーザとの対話における会話エージェントの受話者推定の誤り検出と訂正を実現するために会話中のユーザの反応を用いた誤り検知のための特徴量を抽出した．WOZ 実験を行い，複数人ユーザとエージェントのコーパス収集をした．表情の変化が起きた部分のラベリングを行い，顔認識の結果からその部分を切りだすことで特徴量を求めた．求めた特徴量を機械学習することにより，表情を 90.9% で認識できるようになった．

本研究は，表情機能と音声情報をあわせることで検知をすることを想定しているので，今後の課題として，引き続き音声情報を解析する．また，顔認識ができないときの表情の推定も合わせて検討したい．そして，誤りが起きやすい状況を考慮することにより，誤り検知の可能性を広げたい．さらに，多くの人の表情に対応できるようにするために追加実験も検討する．これらを用い，誤り検知と修復の機能を取り入れた会話エージェントの開発を行う．

参考文献

[1] Huang, H.-H. et al.: Making Virtual Conversational Agent Aware of the Addressee of Users ’ Utterances in Multi-user Conversation using Nonverbal Information, *the Proceedings of the 13th international conference on multimodal interfaces (ICMI ’11)*, pp. 401–408 (2011)

[2] Frampton, M.: et al.Who is you?: combining linguistic and gaze features to resolve second-person references in dialogue, *in the 12th Conference of the European Chapter of the ACL*, pp. 273–281 (2009)

[3] Katzenmaier, M., R. Stiefelhagen, and T. Schultz.: Identifying the addressee in human-human-robot interactions based on head pose and speech, *in international Conference on Multi-modal interfaces*, pp. 144–151 (2004)

[4] Dan Bohus, Eric Horvitz,: Facilitating multiparty dialog with gaze, gesture, and speech *ICMI-MLMI ’10 International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction* Article No. 5 (2010)

[5] 中島慶，藤江真也，松坂要佐，小林哲則: 対話調整的役割を果たす顔表情の認識, 電子情報通信学会技術研究報告 PRMU, パターン認識・メディア理解 105(375), pp. 7–12, (2005)

[6] P.Ekman, W.V.Friesen: Unmasking The Face, Prince Hall, New Jersey, 1975.

[7] 大塚和弘: ノンバーバル行動の観測に基づく会話構造の確率的推論，*VNV* 研究会，(2007)