2H1-4

複数人ユーザ会話におけるエージェントの割り込みタイミングの推定手法の提案

Finding the Timings for a Virtual Conversational Agent to Intervene User-user Conversation

乙木 翔地 Shochi Otogi　堀田 怜 Ryo Hotta　黄 宏軒 Hung-Hsuan Huang　川越 恭二 Kyoji Kawagoe

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

In these days, there are more and more opportunities to use conversational agent as the user interface of the systems in public facilities. However, in those situations, the system is often used by groups of visitors rather than individuals. In the multi-user situation which is much more sophisiticated than single user one, specific features are required. One of them is the ability to intervene user-user conversation. If human agent can do that, it can then engage in more natural conversation, i.e. mixed initiative conversations which occur frequently in human-human conversations. In this study, we aim to develop the method for estimating the appropriate timings for the agent to do intervention. Firstly, we conducted a WOZ experiment for data collection. From the experiment, we found that there can be eight kinds of timings allowing the agent to do intervention. Next, we retrieved the data of face and body directions and acoustic information from the corpus to develope the method for detecting the timings automatically. We also propose the directions of the implementation of a fully autonomous system.

1. はじめに

近年，盛んに会話エージェントが研究されており，今後，ガイドエージェント等の情報提供端末としての発展が期待される．将来，エージェントが美術館や映画館等の公共施設に設置され，職員の代わりに説明や案内を行う場合には，個人での利用だけでなく家族や友達等の複数人のグループで利用する場面が想定される．そういった複数人で利用するエージェントでは，個人で利用するものとは別に，複数人でのインタラクションのために特化した機能が必要となる [1]．我々はこれまでもエージェントと複数人ユーザの会話に関する研究を行なってきた．現在のところ，ユーザからの問いかけに適切に応答するために，受話者推定手法を提案し，リアルタイムシステムの構築を行っている [2].

従来のエージェントと人間とのインタラクションの対話システムでは，ユーザ主導型，もしくはエージェント主導型の 2 種類で成り立っていることが多い．しかし，人と人の対話場面においては互いに割り込みが生じ，主導権が会話の中で移り変わる混合主導型の会話が自然である．複数人ユーザに対しての擬人化エージェントによる情報提供の会話場面において，エージェントがユーザ同士の会話に適切なタイミングで介入し，積極的に有用な情報を提供することが可能となれば，情報提供端末としてより有効なサービスが期待される．これを実現するためには，言語情報，非言語情報の双方が必要不可欠となるが，本研究では，言語情報に対して個人差が少ないと考えられる非言語情報に着目し，会話への割り込みタイミングの推定手法を提案する．非言語情報は対話の場面において重要な要素であると知られている [3, 4]．特に視線や身体の方向は会話参与者の状態を推定するために大きな手がかりになると知られている [5, 6]．まず，本研究ではコーパス収集 WOZ(Wizard of Oz) 実験を行い，どのような場面で割り込み可能であるかを分析する．さらに，ユーザの状態を，発話情況，顔向き，姿勢を用いて推定することで，割り込みタイミングの自動検知を実現する．

Contact: Shochi Otogi, Graduate School of Information Science and Engineering, Ritsumeikan University.
E-mail: is0010hr@ed.ritsumei.ac.jp

2. 関連研究

複数人会話における発話に関する研究はこれまでも行われている．徳永ら [7] は，3 人の女性が「修学旅行に行くならどこがいいか」,「安楽死の是非」,「同性同士の結婚について」をテーマとした会話映像を基に，話者交替以前に発話志向態度を表出し，その態度が他者に解釈されると話者交替に機能すること，また視線方向が発話，非発話の行動調整に寄与することに着目し, 定量的，定性的に分析している．また，河添ら [8] は，エージェントが店員となり商品を販売する会話タスクの中で発話タイミングに影響を与える要因として「発話内容」,「相手との関係」,「発話者の性格」を挙げ，発話タイミングの決定プロセスとしてどれだけ発話を行いたいかを表す発話レベルの定義を行っている．発話レベルは,「話題」,「話の重要度」，相手との「社会関係」,「親密度」から決定され，これをもとに発話タイミングを決定する手法を提案している．しかし，これらの手法は，システムからユーザに積極的に働きかけを行っているものではない．そのため，積極的にユーザに対して働きかけを行うための手法が必要となる．

3. コーパス収集実験

複数人ユーザとエージェントの割り込みタイミングを探るためのデータ収集 WOZ 実験を実施した．

3.1 実験概要

実験環境は，擬人化エージェントが一般的に使用される場面を想定し，エージェントがガイドとなりユーザに対して意思決定の情報提供支援を行う形とした．実験の内容は，スクリーンに投影された等身大の女性のキャラクタの前に 1.8m ほど離れて 2 人 1 組の実験参加者が並んで立ち，対話実験を行った．各グループには，3 つの会話タスクをそれぞれ約 10 分間ずつ取り組んでもらった．対話制御を行う WOZ 操作者はエージェントの反応時間を短縮するために，2 時間操作の練習を事前に行った．WOZ システムには，メニューによる発話選択に加

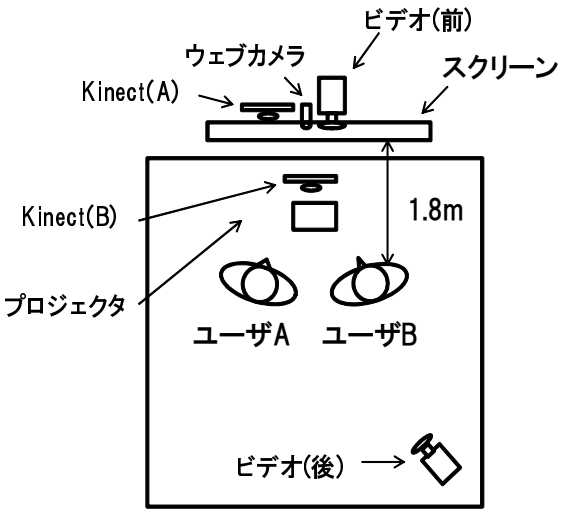

図 1: 実験環境のレイアウト図

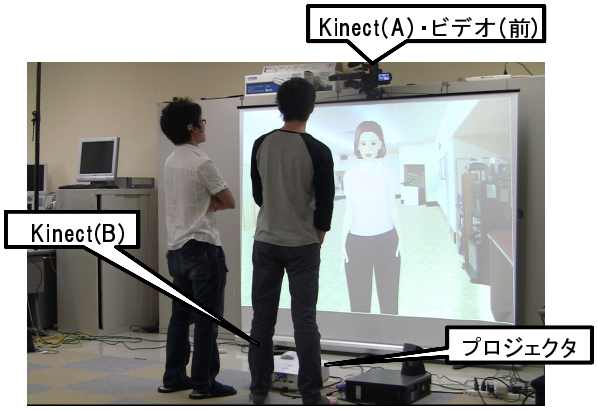

図 2: 実験風景

え，必要な場合に任意の発話をキーボードから入力できる機能も用意した．また，別室の WOZ 操作者が被験者の様子を観察できるように，スクリーンの上に設置したウェブカメラの映像をテレビ電話ソフト Skype で接続した．

3.2 会話タスク

会話タスクは，2 人のユーザがセッション毎のタスクに対してエージェントから提供される選択肢から希望を共同で決定してもらう形とした．ユーザには，前もって選択肢の詳細な事前知識は持っておらず，エージェントに質問をすることにより意思決定のための情報を得ることができる．ユーザ間の活発な議論を誘発するために，最終的に第 1 希望から第 3 希望までを決定することとした．会話は，2 人のユーザが希望の選択肢を決定するまで話し合いをしてもらった．また，実験参加者は大学生・大学院生であったため，大学生に身近なタスクを設定した．それぞれのタスクの概要を以下に示す．

履修登録

実験協力者には，12 の授業のうち一緒に履修したい科目を選択するよう教示をした．協力者は，チュータの役割をするエージェントより，授業に関する情報を得ることができる．エージェントから得られる情報には，授業の概要，担当の先生，単位取得の難易度，授業の時間帯を設定した．

アルバイト紹介

全 14 種類のアルバイトから，どのアルバイトが一緒にしたいかを決めてもらった．エージェントより，時給，勤務地，勤務時間，仕事内容の情報を得ることができ，それぞれの条件から，今の自分と照らし合わせてどのアルバイトがしたいかを決定する．より現実に近い状況にするために勤務地は，大学の周辺を設定し，アルバイト内容は一般的に大学生が行なっているコンビニエンスストアでの販売等を設定した．

観光案内

実験協力者には，1 泊 2 日の九州旅行への国内旅行を行う設定で，訪問する場所を選択するよう教示した．エージェントから得られる情報は，各観光スポットについての簡単な歴史，地元の有名な食べ物，見どころ，観光地である．より実験協力者に興味を持ってもらうために，大学所在地から距離があり馴染みがないと考えられる九州の観光地等を設定した．

表 1: 1 セッションあたりのエージェント・ユーザの発話数

	エージェント	ユーザ	合計
平均	48.0	116.9	164.9
最大	72.0	182.0	254.0
最小	34.0	62.0	96.0

3.3 実験環境

エージェントと被験者とのインタラクションは，図 1 の実験環境のレイアウトのように前方と後方からの 2 台のビデオカメラにより記録した．またビデオデータに加えて，顔方向を取得するためにスクリーン上部に顔認識用の USB ウェブカメラ，姿勢情報取得のための Kinect(A) を設置した．また，被験者の前にも音源情報取得のために 1 台 Kinect(B) を設置した (図 2)．

4. コーパス分析

4.1 会話コーパスデータ

本実験の被験者は，同性の友人 2 人 1 組計 12 組，24 人の学部生，大学院生に協力をしてもらった．12 組のうち，男性は 8 組，女性は 4 組である．平均年齢は，19.2 歳であった．各組には，上記の 3 つのタスクの会話をそれぞれ 1 回ずつ行なってもらった．1 度目は，エージェントの対話に慣れてもらうための練習とし，2 回目以降のデータを 1 組 1 セッションの計 12 セッションを扱った．会話タスクは、3 つのタスクをそれぞれ 4 つずつとし，カウンターバランスをとった．1 グループあたりの会話平均時間は，12 分 43 秒であり，1 発話あたりの平均時間は 2.55 秒であった．また 1 グループ当たりの発話数は表 1 にまとめた．

4.2 割り込みタイミングの定義

コーパス収集実験より，どのような場面で割り込みが可能であるかの分析を行った．非言語情報のみで分類可能である割り込みタイミングを 4 つ，非言語情報に加えて言語情報を使用して分類可能な定義を合わせて計 8 つの種類を定義した．以下に割り込みタイミングの定義を示す．

∗1 http://www.anvil-software.org/
∗2 http://www.cs.waikato.ac.nz/ml/weka/
∗3 http://www.seeingmachines.com/product/faceapi/

I. 停滞の場面 (Stagnation)
会話が停滞している場面
II. 直近のエージェントの発話に対しての更なる支援が必要な場面 (More support)
Not understand:
ユーザエージェントの提供した情報の内容に疑問が生じている場面
Question:
エージェントから提供された情報に関して新たな疑問が生じている場面
Reaction:
エージェントの提供した情報に対してユーザが良い反応を示した場面
Not hear:
ユーザがエージェントの発話がうまく聞き取れなかった場面
III. ユーザへの回想支援が必要な場面 (Recall support)
Forget information:
エージェントから提供された情報を忘れている場面
Forget utterance:
ユーザが会話中にエージェントに対して行った発話の内容を忘れている場面
IV. 質問内容を議論している場面 (Discussion question)
エージェントにどんな内容の質問をしようかユーザ同士が話し合っている場面

4.3 割り込みタイミング数

計 12 グループで割り込みタイミングの総数は 149 タイミングあった．割り込みタイミング数は 1 グループあたり 12.4 回あり，平均約 1 分あたり 1 回の割合でエージェント割り込めることが示された．またグループごとにみると割り込みタイミングが最も多いグループで 28 回あり，最も少ないグループでは 9 回でありばらつきがあった．1 組あたりの割り込みタイミング数を図 3 を示す．

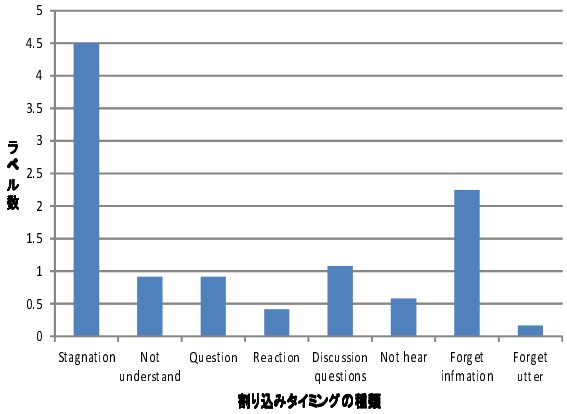

図 3: 1 組あたりの割り込みタイミング数

4.4 ビデオアノテーション分析

分析を行うにあたって取得した音声情報，姿勢，顔方向と割り込みタイミングとの対応を見やすくするためにビデオデータのアノテーション手法を用いた．アノテーションツールには，Anvil[*1]を使用した．定義を行った割り込みタイミング，コーパス収集実験より取得した，音源情報，姿勢情報，顔方向についてアノテーションを行った．以下にそれぞれの分析について示す．

音声データ分析
発話時間・発話者
音声情報では，Kinect のマイクロフォンアレイを用いて取得した発話者，発話時間を基に発話者（エージェント・ユーザ右・左）の計 3 種類のラベルを自動で付与した．また，音声データが均一に収集できない部分があり，取得しきれなかった部分に関しては，手作業で補正した．
受話者
発話者，発話時間のラベリング結果を基に，ユーザの発話が誰に向けられたものかを知るために，受話者について手作業でラベル付を付けた．ラベルは，エージェント・パートナー（もう片方のユーザ）とした，またエージェントの発話はすべてユーザ 2 人を受話者とした．
顔方向分析
顔方向の分析については，頭部方向を顔方向の近似値として利用した．実験時に，ウェブカメラのビデオデータからリアルタイムで顔認識ソフトウェア FaceAPI[*2]を用いて，三次元の頭部の位置と回転角度のデータ及び取得の確信度 (0 または 1) を秒間 30 フレームで取得した．これらのデータをデータマイニングツール Weka[*3]の決定木 J48(C4.5) により分類し，顔方向を正面，パートナー方向，その他，3 種類のラベルを定義し，FaceAPI で取得したログより自動でラベリングを行った．またリアルタイムで顔方向が取得しきれなかった部分に関しては，オフラインでスクリーン上に設置したビデオデータを用いデータの再取得を行い，分類を行った．
姿勢分析
Anvil を用いて手作業で 4 方向（正面，もう片方のユーザに対して 45 度以下の方向転換，45 度以上方向転換，その他）についてラベリングを行った．

5. 割り込みタイミングの自動検知

5.1 自動検知手法

定義を行った非言語情報のみでの分類可能な割り込みタイミングをビデオアノテーションによって可視化した情報を基に自動検知システムの手法を開発した．分類にのために設定した特徴量を以下に示す．また自動検知に使用した各定義の判定条件を表 2 に示す．

発話情報

$U_1(p)$: 発話者 (agent，right_user，left_user)

$U_2(t)$: 各発話間のインターバル (秒)

$U_3(t)$: 無音発話時間 (秒)

$U_4(p)$: 発話に対しての受話者 (agent，right_user，left_user)

$U_5(p)$: 直近の発話に対しての受話者 (agent，right_user，left_user)

$U_6(t,s)$: 一定時間 (t，秒) における発話数 (s，回数)

顔向き情報

$F_1(d)$: ユーザの発話中の顔方向 (agent，partner，others)

$F_2(c)$: 顔向きの方向遷移 (agent → partner，partner → agent)

$F_3(t)$: ユーザの顔の向き合いの継続時間 (秒)

表 2: 各定義の判定条件

定義の種類	判定条件
Stagnation	U_3(t ≧ 4.8) & U_5(p ≠ agent)
More support	U_6(t=8.3, s ≧ 3) & U_2(t ≦ 1.5) & F_3(t ≧ 2.0)
Recall support	U_6(t=10.2, s ≧ 3) & F_3(t ≧ 5.1) & P_3(t ≧ 5.1) & U_2(t ≧ 1.0) & U_1(p ≠ agent) & U_4(p ≠ agent) & U_5(p ≠ agent)
Discussion question	U_6(t=6.3, s ≧ 2) & U_1(p ≠ agent) & U_4(p ≠ agent) & U_2(t ≦ 1.0) & F_2(agent → partner) & P_2(agent → partner)

表 3: 自動検知システムの評価結果

分類	再現率	適合率	F 値
Stagnation	–	–	–
More support	0.46	0.46	0.46
Recall support	0.43	0.40	0.42
Discussion question	0.31	0.43	0.36

姿勢情報

$P_1(d)$: ユーザの発話中の姿勢方向 (agent，partner，others)

$P_2(c)$: 姿勢の方向遷移 (agent → partner，partner → agent)

$P_3(t)$: ユーザの姿勢向き合いの継続時間 (秒)

5.2 分類精度と考察

提案した自動検知手法の分類精度を評価するには再現率，適合率，F 値を用いた (表 3).

表 3 より，Stagnation については，一定時間でユーザの発話がない場合には，会話が停滞していると判断し，必ず何らかの支援が必要であると考えた．定義通りに検出できるため，精度の評価ができない．一方，他の 3 つの分類に関しては，F 値平均が 0.4 前後の結果となった．これは，ランダムでタイミングを算出する場合に非言語情報で検知可能な 4 分類の割り込みタイミング，それ以外の割り込みタイミングの全 5 種類のチャンスレベル 0.2 と比較すると 2 倍近い値となっている．これより，自動分類に関しては，まだ十分な精度を達成できていないものの，現在使用している特徴量が割り込みタイミング検出の精度向上に寄与していることがわかった．

6. おわりに

本論文では，複数人会話におけるエージェントからユーザへの積極的に情報提供を行うための割り込みタイミング検知手法の提案を行った．まず，エージェントの割り込みタイミングを推定するために WOZ 実験を行い，会話データを収集した．次に，収集した音声情報，顔方向，姿勢方向について分析を行った．その後，分析結果より割り込みタイミングを 4 分類，計 8 つ定義した．更に，その定義より，自動検出手法の提案を行い，精度評価を行った．評価結果は，F 値平均が 0.4 前後となった．これは，リアルタイムシステムに実装するには十分とは言えない．今後，より検知精度を高めるために，表情の認識，音声情報 (ピッチ，パワー，話速) の変化等の更なる非言語情報の利用が考えられる．また，非言語のみで分類可能な 4 種類だけでなく，定義を行った言語情報も含めて分類可能だと判断した 8 種類について自動検知を可能にし，その後，割り込みタイミングを自動検知する会話エージェントを実装する予定である．

参考文献

[1] David Traum. Issues in multiparty dialogues. In Advances in Agent Communication,International Workshop on Agent Communication Language (ACL '03) 27, pp. 201–211, 2004.

[2] 馬場直哉, 黄宏軒, 中野有紀子. 人対会話エージェントとの多人数会話における頭部方向と音声情報を用いた受話者推定機構. 人工知能学会論文誌, Vol. 28, No. 2, pp. 149–159, 2013.

[3] 横山真男, 青山一美, 菊池英明, 帆足啓一郎, 白井克彦. 人間型ロボットの対話インタフェースにおける発話交替時の非言語情報の制御. 情報処理学会論文誌, Vol. 40, No. 2, pp. 487–496, feb 1999.

[4] Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, Vol. 23, No. 2, pp. 283–292, 1972.

[5] 大塚和弘, 竹前嘉修, 大和淳司, 村瀬洋. 複数人物の対面会話を対象としたマルコフ切替えモデルに基づく会話構造の確率的推論 (ヒューマンインタフェース基礎). 情報処理学会論文誌, Vol. 47, No. 7, pp. 2317–2334, Jul 2006.

[6] Emanuel A Schegloff. Body torque. *Social Research*, pp. 535–596, 1998.

[7] 徳永弘子, 武川直樹, 寺井仁, 湯川将英. 発話志向態度の表出・理解と発話調整に基づく話者交替分析：3 人会話における「話したい/聞きたい」態度表出の効用 (言語コミュニケーションとそのフィールド). 電子情報通信学会技術研究報告. HCS, Vol. 27, pp. 49–54, 2010.

[8] 河添麻衣子, 北村泰彦. 発話タイミングを考慮したマルチエージェント対話システム. 電子情報通信学会技術研究報告. AI, 人工知能と知識処理, Vol. 106, No. 617, pp. 53–56, mar 2007.