2D5-9

ベイジアンネットワークを用いたヒューマンエージェントインタラクションのモデル化

Modeling Human-Agent Interaction using Bayesian Network Technique

中野有紀子[*1] Yukiko Nakano
村田 和義[*1] Kazuyoshi Murata
榎本 美香[*1] Mika Enomoto
有本 泰子[*2] Yoshiko Arimoto
朝 康博[*3] Yasuhiro Asa
佐川浩彦[*3] Hirohiko Sagawa

[*1] 東京農工大学 Tokyo University of Agriculture and Technology
[*2] 東京工科大学 Tokyo University of Technology
[*3] (株) 日立製作所中央研究所 Central Research Laboratory, Hitachi, Ltd.

Task manipulation is direct evidence of understanding, and speakers adjust their utterances in progress by monitoring listener's task manipulation. Aiming at developing animated agents that can control multimodal instruction dialogues by monitoring user's task manipulation, this paper presents a probabilistic model of fine-grained timing dependencies among multimodal communication behaviors. Our preliminary evaluation revealed that our model can predict grounding judgment and user's successful mouse manipulation quite accurately, suggesting that the model is useful in estimating user's understanding, and can be applied to determining the agent next action.

1. はじめに

アプリケーションソフトウェアにおけるヘルプは，システムの使い方がわからなくなった場合に，ユーザを支援してくれる重要な機能であり，システムの使いやすさ，わかりやすさに関するユーザ評価に大きな影響を与えるものである．しかし，効果的なアドバイスでなければかえってユーザの作業を妨げる結果となってしまうため，適切なタイミングで適切なアドバイスを行うヘルプシステムが望まれる．

本稿では，この問題に対する 1 つのアプローチとして，ヘルプ機能をユーザとエージェントとのマルチモーダルなインタラクションとしてとらえ，会話的に操作方法を教示するヘルプエージェントを実現するための基礎となる会話モデルをベイジアンネットワークを用いて構築し，その精度を評価する．

2. 背景

2.1 聞き手のモニタリングと発話の調整

(Clark and Krych 2004) は，作業空間を共有する状況での会話を詳細に分析し，課題遂行方法を指示する話し手が，聞き手の言語的な応答のみならず，ジェスチャーや表情等の非言語的な反応，さらには聞き手による作業遂行動作やそれによって生じる状況の変化に応じて，発話途中で指示の内容を動的に変化させている事例を紹介している．このような，課題遂行状況のモニタリングに基づく発話の調整は，マルチモダリティを利用して会話を効率的に進めるための重要な機能であると考えられる．

ビデオ映像を通してお互いの姿を確認できる状況で，画面を共有しながらソフトウェアの使い方を指示している会話の例を**図 1**に示す．ここでは，話し手は，「番組検索ボタン」の位置を指示しているが，ポーズをさしはさんで，聞き手のマウスの動きの様子を見ながら，こまぎれに指示を与えている．また，ジェスチャーストロークのタイミングも調整されている．このように，聞き手の様子に合わせて言語的・非言語的行動を調整することにより，聞き手からの言語的な応答が全くないにもかかわらず，円滑に会話を進めているように思われる．

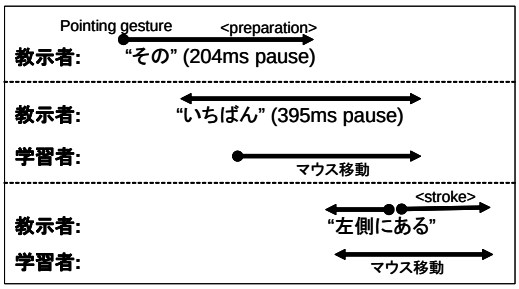

図 1: 課題操作場面での会話例

2.2 非言語情報による発話のグラウンディング

このような対話をユーザとエージェントの間で実現するためには，現在のユーザの作業状況から，課題や対話の状態を予測し，ヘルプの内容を変更する機構が必要となる．例えば，ユーザがエージェントの指示を十分に理解してないようであれば，説明を追加すべきであるし，ユーザの理解が十分であり，指示した操作が近い将来実行されると予想されれば，余計な指示は行わず，ユーザからの入力を待つほうが望ましいと考えられる．

(Paek and Horvitz 1999)は，音声対話システムに，ベイジアンネットワークを用いた対話モデルを適用し，確率推論により，音声対話における相互理解の不確実性を考慮した対話制御を実現している．さらに(Horvitz, Breese et al. 1998)では，同様の手法をヘルプシステムにおけるユーザモデリングに利用している．

そこで本稿では，このような確率推論の手法をマルチモーダル対話のモデル化に利用し，課題の遂行状況や発話の基盤化(grounding)(Clark and Schaefer 1989)の判断の精度を評価する．

連絡先：中野有紀子，東京農工大学大学院工学府情報工学専攻，〒184-8588 東京都小金井市中町 2-24-16，e-mail: nakano@cc.tuat.ac.jp

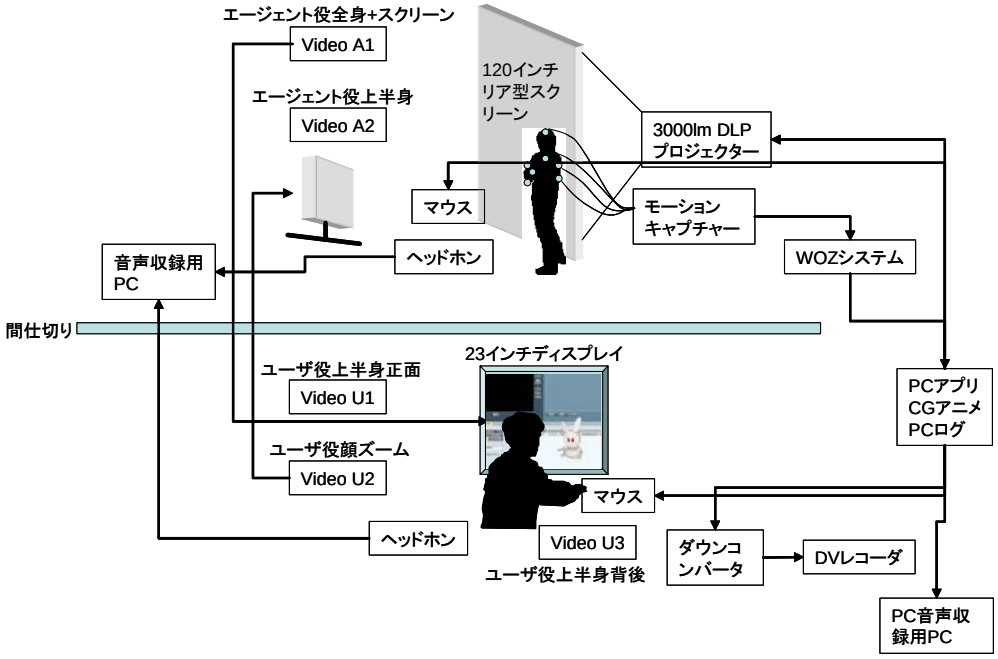

図 2: 会話データ収録環境

3. 会話データの収集とコーパスの作成

本節で, 対話モデルの構築のためのコーパス作成について述べる. まず, 対話データ収集のために, テレビパソコンの操作を課題とし, デスクトップに表示されるアニメーションキャラクターを用いた Wizard-of-Oz 法による会話収録実験を行った.

3.1 収録方法

ヘルプエージェント役の被験者(「エージェント役」と呼ぶ)とユーザ役の被験者(「ユーザ役」と呼ぶ)はそれぞれ別室で実験に参加する. 収録環境を**図 2**に示す. 操作対象となる PC の出力はユーザ役が着席している机に置かれた 23 インチのモニターに映し出されると同時に, エージェント役の背後に設置された 120 インチの大型スクリーンにも投影される. エージェント役はスクリーンに映し出された PC 出力と, スクリーンの正面に設置された小型ディスプレイに映るユーザ役の顔映像(Video U2)を見ながら指示を行う.

また, エージェント役の上半身には 10 個のモーションセンサーが装着され(**図 3**(a)), キャプチャーされたエージェント役の動作データは Wizard-of-Oz システムに送られ, ウサギ型の CG キャラクターアニメーションを制御する. さらに, このキャラクターアニメーションは操作 PC の画面に合成される. 従って, ユーザ役のモニターおよびエージェント役のスクリーンには**図 3** (b) のような画面が映し出される.

(a)エージェント役被験者　　(b)PC 画面

図 3 : エージェント役被験者と PC 画面

エージェント役およびユーザ役の被験者は共にヘッドセットを装着し, それを通して会話を行う. ただし, エージェント役の音声は音声変換ソフトウェア, Herium により機械的な音声に変換された. 両者の音声はヘッドセットのマイクロフォンから収録され, USB オーディオキャプチャーを通して, WindowsPC 上に wav 形式で保存された.

3.2 課題と実験デザイン

テレビパソコンの操作課題として番組録画の課題と, DVD への書き込み操作の課題がそれぞれ 2 種類ずつ, 計 4 種類用意され, どちらか一方を実施することとし, 条件の割り振りに偏りがないよう被験者ごとにカウンターバランスをとった.

エージェント役 10 人, ユーザ役 20 人, 計 30 人の被験者が実験に参加した. 各エージェント役が相手を変えて 2 セッションずつ行うことにより, 計 20 ペアの対話を収録した.

4. コーパスの作成

200 ミリ秒以上のポーズが検出されたところで, 音声データを区切り, これを IPU (Inter Pausal Unit)とし, 書きおこしの単位とした. 書き起こされた対話データのうち 25 対話について以下のタグを付与することにより, コーパスを作成した.

4.1 発話内容タグ

課題の特徴に着目し, 各 IPU で言及されている内容について, 以下のような分類を行った.

- Identification(id) : 操作対象物を指定する発話
- Operation(op) : 操作の実行を指示する発話
- Identification + Operation(idop) : 1 つの IPU 内で id と op を行う発話
- State(sta) : 操作前もしくは操作後のシステム状態について言及する発話

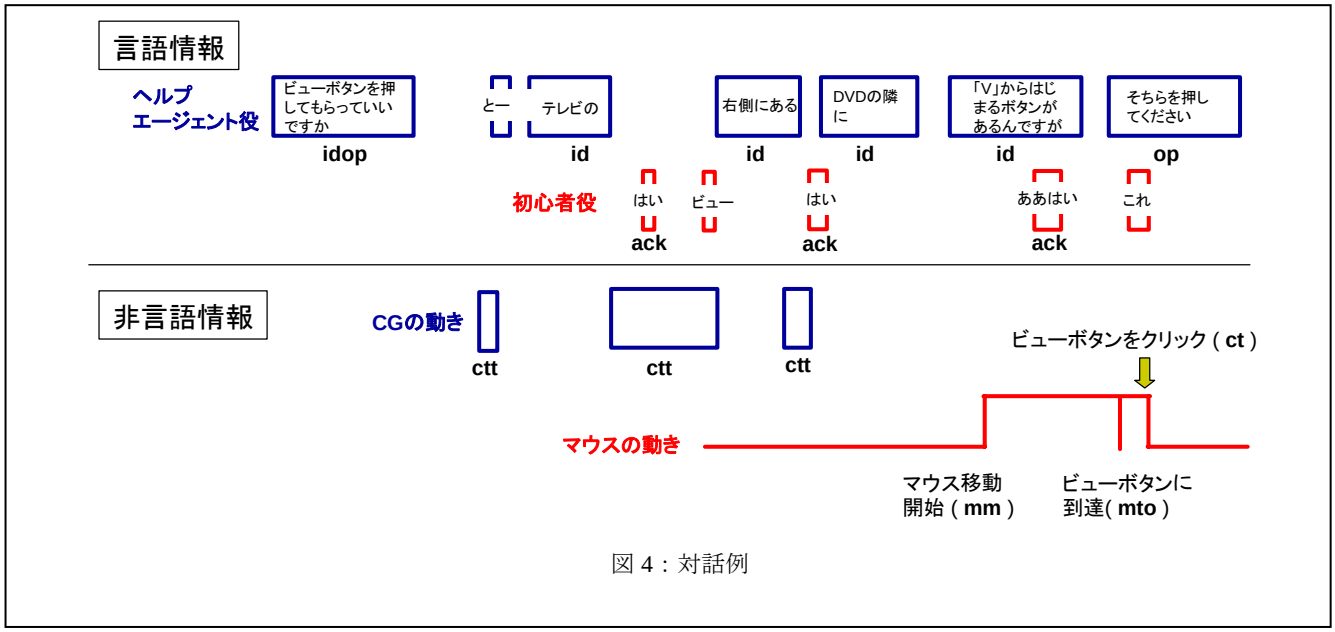

図 4：対話例

- Function（fn）：アイコンやシステムの機能を説明する発話
- Goal（gl）：操作の最終的な目的や内容を決定するための発話
- Acknowledgement（ack）：合意や相槌の発話

以上の発話内容タグに関する 2 名の作業者間での一致率は K=0.89 であり，高い一致率が得られた．

4.2 ジェスチャーとキャラクターの動き

(1) エージェント役のジェスチャー

ジェスチャーの形態とフェーズについて次のようなタグを付与した．

(1-1) ジェスチャーの形態

- ポインティング：モニターのある場所を指し示す動作．円を描くようにして、複数の対象を囲うような動作もポインティングとみなす．
- トレース：画面中の文字をなぞるようにして指し示すジェスチャー
- その他：上記以外のジェスチャー

(1-2) ジェスチャーのフェーズ

- Preparation：ジェスチャー開始から stroke までの腕の動き
- stroke:ジェスチャーの最高潮の部分．指差した瞬間や、腕を振ったその瞬間など
- beats:強調のために stroke を繰り返す動き
- hold:stroke 状態での停止．ポインティングである 1 点をさし続けている場合など
- retract:stroke の位置から腕を戻す動作
- partial-retract:ある stroke から次の stroke に移行するため腕を戻す動作
- hesitate:partial-retract と preparation の間の休止．stroke から stroke への移動の間の停止期間

(2) CG キャラクターの動き

CG キャラクターの位置とジェスチャーに関し，以下のタグを付与した．

(2-1) CG キャラクターの移動：CG キャラクターの位置が変化している区間．15 フレーム以上静止していた場合移動が終了したとみなす．

(2-2) CG キャラクターによるポインティング： CG キャラクタの腕部分が操作対象物に接触している区間．

4.3 マウス操作

ログ自動収集プログラムを用い，ユーザ役による以下の 3 種類のマウス操作のログを収集した．

- マウスの移動
- マウスクリック
- マウスが置かれている対象（ボタン名，メニュー名等）

これら自動的に収集されたマウスログデータの不具合を多少修正することにより，マウス移動が生じた時間，マウスクリックが起こった時間，マウスが置かれている対象について Anvil 形式 (Kipp 2004)のアノテーションデータを作成した．さらに，マウスの滞留位置が現在の対話の焦点となっている対象物上である場合には，その情報をマウス操作のタグに追記した．

4.4 会話データ例

図 4に収集した対話の一例を示す．エージェント役が「ビューボタンを押してもらっていいですか」と指示を出すが，これに対してユーザ役が全く反応を示さないので，それ以降エージェント役は説明の方略を変更し，ボタンの位置についての説明をポーズをはさんだ短い発話片を用いて行っている．これに対し，ユーザ役は適宜あいづちを返しているものの，依然としてマウスの動きが観察されないので，ユーザ役はさらに説明を付加していく．「V からはじまるボタンがあるんですが」という発話の途中で，ユーザ役のマウスが移動しはじめ，ビューボタンに到達したことを確認すると，ユーザ役は，その次の発話で「そちらを押してください」という動作の指示を述べている．このように，エージェント役はユーザ役による課題遂行の状況をモニタしつつ，指示内容やそのタイミングを調整していることがわかる．

5. ベイジアンネットワークによる対話のモデル化

ベイジアンネットワークの手法を利用し，作成したコーパスからユーザ役対ヘルプエージェント（エージェント役）の統計的対話モデルを構築した．

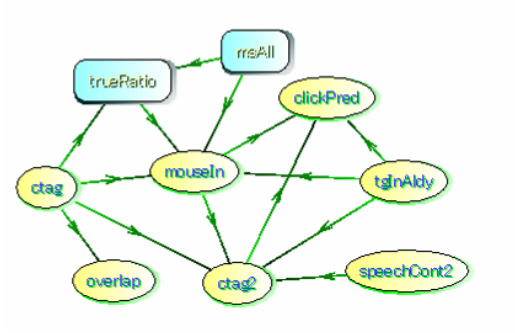

図 5：ベイジアンネットワークモデル

まず，コーパスをエージェント役の各発話の開始時間から 500 ミリ秒毎に区切り，全部で 1246 個のデータ区間を得た．この中からランダムに 124 個をテストデータとし，残りの 1122 個を学習用データとした．モデルの構築には，過去 1 秒間の履歴，現在のデータ区間の情報，数区間先の状態に関して計 9 つのノードを設定し，統計的学習を実行した．構築したベイジアンネットワークモデルを**図 5**に示す．

ここでは，(1) 過去 2 区間(1 秒間)におけるエージェントのポインティングジェスチャー継続時間の割合，(2) 過去 2 区間におけるユーザのマウス移動の回数，(3) 現 IPU の内容タグの種類，(4) 現区間の終了時に現 IPU が終了するか否か（現区間の終了直後に発話ポーズが後続するか），(5) マウスが操作対象のボタンやメニュの上にあるか否か，の 5 種類の情報をネットワークに与え，発話基盤化 (grounding) の予測とマウスクリックの予測を行った．その予測精度を表 1 に示す．

表 1: 予測精度

	ベースライン	提案モデル	精度向上
発話基盤化の判断	0.78	0.88	10%
マウスクリック予測	0.78	0.88	10%

5.1 発話基盤化の予測

エージェント役が次の IPU で説明を補足するか，次の説明に進むか，つまり現在の発話が基盤化されるか否かを予測した．閾値は精度を最大化するように設定された．全ての場合に対して基盤化されないと判断した場合(ベースライン)の精度が 78%であったのに対し，作成したネットワークモデルを用いて予想した場合，精度は 88%であり，10%の向上が見られた．

5.2 マウスクリックの予測

構築したベイジアンネットワークを用いたマウスクリックの予測として，ユーザが近い将来（5 区間，つまり 2.5 秒以内）に操作対象に対して正しいマウスクリックを実行するか否かを予測した．その結果，ベースラインが 78%であったのに対し，本モデルではクリックの予測精度は 88%であった．

6. おわりに

本稿では，会話的なヘルプエージェントの実現を目指し，その基礎データとなるコーパスの構築とベイジアンネットワークを用いたマルチモーダル対話モデルについて報告した．

今後は，さらにコーパスの整備を進め，マルチモーダ対話に関するより詳細な統計的分析を行う．また，複数のモダリティの相互依存関係を統合的にモデル化する方法について検討を進め，マルチモーダルな対話の co-constructive な構築プロセスを解明するとともに，それをヘルプエージェントの会話機構として実装していく予定である．

参考文献

Clark, H. H. and M. A. Krych (2004). "Speaking while monitoring addressees for understanding." Journal of Memory and Language **50**(1): 62-81.

Clark, H. H. and E. F. Schaefer (1989). "Contributing to discourse." Cognitive Science **13,**: 259-294.

Horvitz, E., J. Breese, et al. (1998). The Lumiere Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users. Fourteenth Conference on Uncertainty in Artificial Intelligence.

Kipp, M. (2004). Gesture Generation by Imitation - From Human Behavior to Computer Character Animation, Boca Raton, Florida: Dissertation.com.

Paek, T. and E. Horvitz (1999). Uncertainty, Utility, and Misunderstanding: A Decision-Theoretic Perspective on Grounding in Conversational Systems. Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems. S. E. Brennan, A. Giboin and D. Traum. Menlo Park, California, American Association for Artificial Intelligence**:** 85-92.