1K3-OS-17a-5

多人数対話ロボットの実現にむけたマルチモーダル対話データの収集と分析

Collection and analysis of multi-modal dialogue data for communication robot with multi-participant

石川 真也[*1] 船越 孝太郎[*2] 篠田 浩一[*3] 中野 幹生[*4]
Shinya Ishikawa Kotaro Funakoshi Koichi Shinoda Mikio Nakano

[*1*3]東京工業大学 大学院情報理工学研究科 計算工学専攻
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology

[*2*4](株) ホンダ・リサーチ・インスティチュート・ジャパン
Honda Research Institute Japan Co., Ltd.

For obtaining a natural interaction between Human and Robot, Robot need to classiry varoius dialogue states such as who are participating, which participant talks to whom, and whether a participant understands Robot's speech or not. The purpose of our research is to construct a multi-modal dialogue system which can communicate with multiple people simlutaneously. We collected multi-modal data where Robot operated by a human in the WOZ environment interacts with three participants. We annotated the data with several tags such as speech addressee, gaze, and how the person perticipate in the conversation. In this paper we explain the recording environment, summarize the data description, and report the result of a preliminary experiment of speaker identification where we do not use the raw data, but use the tag information.

1. はじめに

実環境で動作し, 人間と共存するロボットの実現に向けた研究が行われている. 特にロボットが複数の人間と同時にコミュニケーションを取るような状況においては, 単に音声認識によって話している内容を獲得するだけではなく, 「誰が会話に参加しているか」「誰が誰に話しかけているか」などの状態を適切に識別する能力が必要である. 本稿では, 主に話者の発話の対象を中心とした, 識別すべき状態を下記によって規定し, これを対話状態と呼ぶ.

発話対象 話者が誰に向かって話しかけているか

注視対象 話者がどこを注視しているか

参加状態 話者がどのように会話に参加しているか

人間同士の対話において, 対話状態を識別するためには, 人間の視線遷移やジェスチャーなどの非言語情報が有用である.

ロボットと複数人の対話において, 非言語情報から対話状態を機械的に推定する研究は複数報告されている. 中島らは, 机の上に配置された 2 体のロボットに搭載されたマイクとカメラから得られる音源定位結果と顔検出結果に基づき, アクティブに会話に参加しているユーザの推定をおこなっている [中島 13]. しかし, 中島らの設定したドメインは机を囲んでの会話であり, 会話参加者の移動や入退出は考慮されていない. また, Vinyals らは, 音声対話システムに入力された動画像と音声, またそれらから抽出される種々の非言語情報を入力として, 統計的手法によって音声区間と音声重畳区間の検出や, 発話者, 発話対象の推定をおこなっている [Vinyals 12]. このとき Vinyals らは, モニタ上に投影されたバーチャルエージェントを用いてデータを収集している. Jayagopi らは, ロボットがクイズを出題し, 2 名の話者がそれに答える会話を収録し, 話者の視線と発話履歴, クイズの難易度を特徴量として用いて発話対象の推定実験をしている [Jayagopi 13]. このとき会話への参加者の数は 2 名に固定されており, 話者の発話対象はロボットかもう一方の話者の二者択一としている. そのため, 会話への参加者の数が増減する環境における発話対象推定に関しては検討されていない.

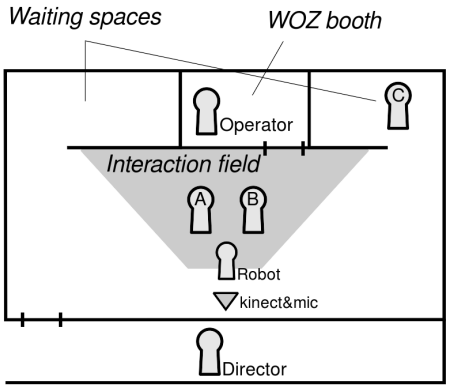

図 1: データ収録環境見取り図

本研究は, ロボットが複数の人間と同時に会話する状況を想定し, 音声に含まれる言語情報と, 視覚情報から抽出される人間の顔向きや表情, ジェスチャーなどの非言語情報から, 対話状態を推定するシステムの構築を目的とする. そして, 本稿ではシステムの構築に先立ち, Wizard of OZ 法を用いてロボットと複数の人間の対話を収録し, それに対話状態を示すタグを付与することで構築したマルチモーダル対話コーパスについて報告する. 以下で構築したコーパスの概要と, その分析結果, そして今後の課題を述べる.

連絡先: 石川 真也, 東京工業大学, ishikawa@ks.cs.titech.ac.jp

2. データ収録

互いに知人関係にある 3 名を 1 組とし, 30 組 90 名の参加者を用いて, マルチモーダル対話データを収録した. 各組はロボットと 25 分のセッションを 2 回おこなう. 収録したデータの総数は 60 セッション, 1500 分であった. ロボットは Aldebaran 社製の人型ロボット「NAO」[Ald] を用いた. 収録環境の見取り図を図 1 に示す. 収録は 3 名の被験者（それぞれ A, B, C とする）と, ロボットを操作するオペレーター, 被験者に指示を与える監督者の 5 名によっておこなわれる. 収録環境は被験者がロボットと会話をおこなうインタラクションフィールド（以下フィールド）, その両端にある衝立で仕切られた待機スペース, オペレーターが入る WOZ ブースと, 監督スペースから構成される. 被験者の顔向きや視線, 立ち位置の非言語情報を収録するために, 動画の収録には Microsoft 社の Kinect を用いて, 毎秒 30 フレームの RGB 画像と深度画像を収録した. また, 音源方向を獲得するために, ピンマイク 4 本をロボットの後ろに等間隔に配置し, 4 チャンネルの音声を収録した. インタラクション中のロボットの発話やジェスチャーについて, タイムスタンプ付きのログを記録した. 以下では監督, 被験者とオペレーターそれぞれの役割と収録の流れについて説明する.

監督者は壁で隔たれた監督スペースから, あらかじめ定めたシナリオに則って, 被験者に無線で入退場の指示を出す. 監督者が被験者に与える指示は下記のいずれかである.

- フィールドに入って, ロボットとの会話に参加する
- フィールドに入るが, 会話には参加せず, ロボットと他の参加者の会話を傍観する
- ロボットとの会話から離脱し, フィールドから出る
- フィールドを素通りし, 反対側の待機スペースへ移動する

無線のチャンネルは被験者ごとに違うため, 各被験者は, 他の被験者が監督者からいつどのような指示を受けたかを知ることができない.

オペレーターは, 被験者から隔たれた WOZ ブース内からロボットを操作する. オペレーターはロボットに搭載されたカメラからの動画像とロボットの後ろに設置されたマイクからの音声を参照しながら, 制御用アプリケーションを通して, ロボットの顔の向きを変えたり, 定型文の発話やジェスチャー動作をさせる. オペレーターは下記の行動指針に基いてロボットを操作する.

- フィールドに新たな被験者が入場したとき, その被験者のほうへ顔を向け, ゲームへ誘う発話を行う
  - 被験者が誘いを承諾したとき, その被験者を参加者とみなす
    - 被験者がセッション中で初めて会話に参加する場合, 自己紹介発話をする
  - 被験者が誘いを拒否するか, 反応が無ければ, その被験者を傍観者とみなす
- ゲームのシナリオに則って, 参加者に向けて定型文発話やジェスチャーをする
  - 傍観者がフィールドにいる場合, 一定時間経過する度にその傍観者をゲームに誘う

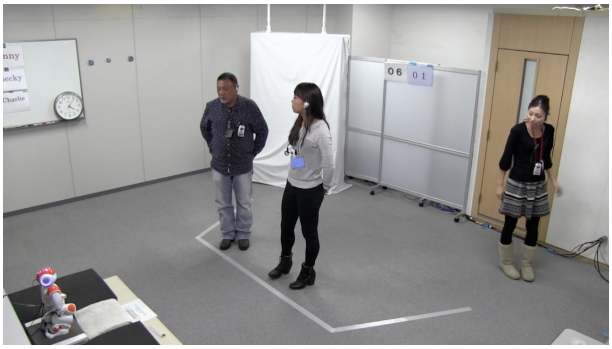

図 2: データ収録風景

- 会話参加者がフィールドから去ろうとしたとき, その被験者のほうへ顔を向け, 呼び止める発話を行う

また, オペレーターはセッションごとに, 一度会話をした被験者の顔と名前を覚える. なお, ロボットが手動で操作されていることは被験者には知らされない.

被験者 3 名は監督の指示に従ってフィールドに入退場し, 図 2 に示すように, ロボットと対面してゲームをおこなう. 被験者はそれぞれ個別に会話への参加と離脱を繰り返すため, ロボット（オペレーター）側から見れば, 会話の参加者の数が 1 名から 3 名まで増減する環境となる. 被験者は日本語と英語の両方で発話できる. 設定上, ロボットの発話は原則全て英語であるが, ロボットの発話が通じずに会話が滞ってしまう場合のみ日本語を使う. 被験者がロボットの発話を聞き取れなかった場合は, 直前の発話内容をロボットに聞き返すことができる. また, ゲーム中, フィールドにいる被験者同士は自由に会話できるため, 被験者同士の会話によってロボットの発話内容や意図を推測することもできる.

被験者の発する非言語情報が会話のドメインに依存するかどうかを調べるため, 2 種類のゲームを用意した. どちらもロボットがホスト役, 被験者がゲスト役となってゲームを進行する. ひとつは「20 のとびら」(20 doors) と呼ばれるクイズゲームで, ホストが思い浮かべたあるもの・ことに対して, ゲストは Yes-No 形式の質問を複数回おこない, それが何であるか推理する. もうひとつは「ジェスチャーゲーム」(Gesture) で, これは 2 つのステップにわかれている.

1. ホストはゲストに, ある英単語に相当するジェスチャーを教示する
2. ホストは英単語を発話するのみで, ジェスチャーをしない. ゲストはステップ 1 で覚えたジェスチャーを思い出して行う. 複数のゲストがゲームに参加している場合, 最も早くジェスチャーができた人を勝ちとする.

ロボットは新しい参加者に対してまずステップ 1 を実行し, 適当な数のジェスチャーを参加者が覚えたと判断したら, ステップ 2 に移行する.

3. コーパス構築

収録したデータに対し, ELAN[Auer 10] を用いて被験者の発話対象, 注視対象, 会話への参加状態に対応する注釈層をそれぞれ設定し, タグを付与した. 注釈層とタグの一覧を表 1 に示す. 下記でその詳細について説明する.

注釈層の名前	注釈層の説明	タグの種類
Participating	会話への参加状態	Participating Observing Passing Leaving
Speech	発話対象	toA, toB, toC, toRobot Monologue Laugter
Gaze	注視対象	toA, toB, toC, toRobot toOthers Invalid

表 1: 注釈層の一覧

図 3: Kinect による録画

Speech 注釈層は被験者の発話対象を示し, そのタグは被験者の一つの発話に対して一つ付与される. 例えば被験者 A が他の被験者 B にむけて発話した場合, その発話には toB というタグが付与される. また, 発話は複数の発話対象を取ることができる. 例えば被験者 A が被験者 B と C に向けて発話した場合, 発話には toBC というタグが付与される. そして独り言に対して Monologue, 笑い声に対して Laughter タグを付与する.

Gaze 注釈層のタグは当該被験者が注視している対象を示し, 被験者がインタラクションフィールドにいるすべての動画フレームに付与される. 具体的には, 図 3 に示すような kinect で取得した画像を参照しながらタグを付与する. 例えばあるフレームにおいて被験者 A がロボットを見ている場合, toRobot というタグが付与される. また, 被験者やロボット以外を注視している場合は toOthers, 眼が隠れたり後ろを向くなどして, 被験者がどこを見ているか判別できない場合は Invalid タグを付与する.

Participating 注釈層は被験者が会話に参加しているか否かを示している. これも Gaze と同様に, 被験者がインタラクションフィールドにいるすべての動画フレームに対して付与される. 被験者がロボットとの会話に参加している場合は Participating, 会話に参加せず, ロボットと他の被験者のやり取りを見ているだけの場合は Observing, インタラクションフィールドを通り過ぎているときは Passing, 会話から離脱しようとしている場合は Leaving タグを付与する.

	被験者	会話参加時間	総発話数
20doors	A	820	115
	B	644	118
	C	747	134
Gesture	A	717	161
	B	672	108
	C	747	140

表 2: 各ドメイン・各被験者の会話参加時間（秒）と総発話数

	参加者数	toRobot	toA,toB,toC	Monologue
20doors	一人	108 (80.0)	8 (5.9)	19 (14.1)
	複数人	162 (66.4)	51 (20.9)	31 (12.7)
Gesture	一人	146 (93.0)	1 (0.6)	10 (6.4)
	複数人	124 (65.3)	51 (26.8)	15 (7.9)

表 3: 会話参加者の各対象への発話回数. （）内は割合（%）

4. コーパスの分析

現在タグ付けが終了している, 同一の三人組による各ドメイン 1 セッションずつ, 計 2 セッションに対する分析をおこなった. はじめに各被験者が会話に参加した時間と, 会話中の発話総数を表 2 に示す. どちらのゲームにおいても, 会話の参加者は 1 分あたり平均 8〜13 回程度の発話をおこなっており, 活発な会話が実現できている.

4.1 ドメインによる差異

付与したタグの内訳と, ドメインによる違いについて分析する. ロボットとの会話への参加者が一人である場合と, 複数名である場合それぞれについて, Laughter を除く発話対象の分布を表 3 に, Invalid を除く注視対象の分布を表 4 に示す. なお, 参加者が一人のときに他の被験者への発話や注視が発生しているのは, Participating 以外の参加状態の被験者に発話や注視をしている場合にあたる.

まずドメインや参加者の数に関わらず, ロボットへの発話や注視が最も多い. これは両ゲームともロボットがホスト役となって進行することを反映している. しかし会話への参加者が二人以上になると, 他の参加者への発話や注視が発生するため, 相対的にロボットへの発話や注視の割合が下がる. ドメイン間での違いについて述べると, 「20 のとびら」では, 他の参加者への発話や注視だけでなく, 参加者が独り言を呟いたり, ロボットや参加者以外の場所を注視する割合が多い. これはロボットに質問を投げかける際に, 自ら現在のクイズの状況を口に出して確認する発話が起きるためである. したがって, ロボットへの発話と独り言を区別する場合には, 独り言を多く含む「20 のとびら」が有用である. また, 「ジェスチャーゲーム」では, ロボットの動きを見て覚えるというルールの影響から, 「20 のとびら」と比較して参加者のロボットへの注視時間が長く, ゲーム中は参加者同士の自由な会話が抑制される傾向がある. ジェスチャーに対応する英単語を参加者が聞き取れない状況が, セッションを通して 13 回発生しており, その場合に参加者同士でロボットの発話内容を推測する発話が起きた. このことから, 人間がロボットの発話を理解できない場面を多く取りたい際は, 「ジェスチャーゲーム」が適していると考えられる.

4.2 注視対象情報を用いた発話対象推定

予備実験として, アノテーションを施した Laughter を除く発話に対して, その発話者と, 発話区間における発話者の注視

	参加者数	toRobot	toA,toB,toC	toOthers
20doors	一人	449 (68.0)	51 (7.7)	161 (24.3)
	複数人	1084 (66.1)	307 (18.7)	249 (15.2)
Gesture	一人	732 (86.9)	49 (5.8)	61 (7.3)
	複数人	1051 (78.9)	137 (10.3)	144 (10.8)

表 4: 会話参加者の各対象への注視時間（秒）．（）内は割合（%）

ルールの種類	20doors	Gesture
AllRobot	72.1	75.5
FirstGaze	68.3	77.1
LongestGaze	83.3	82.1

表 5: ルールベースによる発話対象推定の正解率（%）

対象が正しく獲得できると仮定し, 発話対象を toA, toB, toC, toRobot, Monologue の中から推定した. 即ち人手で付与した Gaze タグの情報を参照し, 下記 3 つのルールそれぞれに従って発話対象を推定し, ルール毎の正解率を比較した.

AllRobot 発話対象を全てロボットとする.

FirstGaze 発話区間の最初に発話者が注視している対象を発話対象とする. 注視対象が他の参加者やロボット以外であるとき, 発話は独り言であるとする.

LongestGaze 発話区間の中で発話者が最も長く注視している対象を発話対象とする. 注視対象が他の参加者やロボット以外であるとき, 発話は独り言であるとする.

タグの付与が終了している 1 組 2 セッション分のデータを用いた結果を図 5 に示す. LongestGaze を用いた推定が最も正解率が高く, 発話対象推定における視線情報利用の有効性を示唆する結果となった.

また, LongestGaze によって推定した発話対象と実際の発話対象が一致しない発話のデータを観察したところ, 例として下記のような特徴的な誤りを発見した.

1. 他の被験者への発話を継続しながらロボットの方へ向き直る場合. このとき発話対象は他の被験者だが, 注視対象はロボットである.
2. ロボットを見ながら独り言をつぶやく場合
3. 被験者が後ろを向く, 両目を手で覆うなどして, 被験者の注視対象が取れない場合

これらの誤りは, 発話者の注視対象以外の特徴量を使うことで解決が期待できる. 例えば誤り 1 に関して, 他の被験者の注視対象や発話者の直近の発話対象を用いることで正しく識別できる可能性がある. 誤り 2 では, 独り言を話す際の声量が, ロボットへの発話のそれと比較して小さいため, 音声のパワーが特徴量として使えると考えられる.

5. 課題

コーパス分析においては, 被験者の注視対象が正しく獲得できる仮定のもとで, 人手で付与した注視対象タグを用いた. 次の課題として, 動画像と深度画像から会話参加者の顔向きを推定することで, 注視対象を精度よく推定できるかを調べる.

また, 今回は対話における発話対象を推定したが, 人が発話内容を理解できていない状態を検知したり [Forbes-Riley 11], 人がロボットとのやりとりに飽きている状態などを検知できれば, それに応じてロボット側から人へ, 会話への参加を促す発話が可能になると考えられる.

さらに, 新たな対話状態や, 新たな特徴量を追加する度に, ドメイン知識に基いて人手でルールを定めることは工数が大きい. この問題に対しては, 機械学習を用いて, 対話状態推定のルールを自動的に獲得する手法 [Vinyals 12] を試行する予定である.

6. おわりに

本稿では, 複数人とロボットが会話するドメインを設定し, Wizard of OZ 法を用いて収録したデータに対して種々の対話状態を示すタグを付与することでコーパスを構築した. ドメインによるデータの差異について分析をおこなった. また, 人手で付与した注視対象の情報を用いて, 被験者の発話対象を推定する実験をおこなった結果, 発話区間内で最も長い時間注視していた対象を発話対象とする場合に最も高精度であった. 今後の課題として, 発話対象推定に用いる, 視線情報以外の特徴量の検討, 発話対象以外の対話状態の推定, 機械学習を用いた状態推定などを挙げた.

参考文献

[Ald] NAO Aldebaran robotics, http://www.aldebaran-robotics.com/

[Auer 10] Auer, E., Russel, A., Sloetjes, H., Wittenburg, P., Schreer, O., Masnieri, S., Schneider, D., and Tschpel, S.: ELAN as Flexible Annotation Framework for Sound and Image Processing Detectors, in *LREC'10*, Valletta, Malta (2010), European Language Resources Association (ELRA)

[Forbes-Riley 11] Forbes-Riley, K. and Litman, D.: Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor, *Speech Communication*, Vol. 53, No. 9-10, pp. 1115–1136 (2011)

[Jayagopi 13] Jayagopi, D. B. and Odobez, J.-M.: Given that, should I respond?: contextual addressee estimation in multi-party human-robot interactions, in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, HRI '13, pp. 147–148, Piscataway, NJ, USA (2013), IEEE Press

[Vinyals 12] Vinyals, O., Bohus, D., and Caruana, R.: Learning Models for Speaker , Addressee and Overlap Detection from Multimodal Streams, *ICMI '12*, pp. 417–424 (2012)

[中島 13] 中島 大一, 駒谷 和範, 佐藤 理史：複数人会話におけるロボットによる視聴覚情報に基づくアクティブユーザの推定, 情報処理学会研究報告 SLP-95, pp. 1–8 (2013)