3C2-OS19-7

思考喚起型対話におけるユーザ対話意欲の分析

Analysis of user engagement in thought-evoking human-robot dialogue

堂坂浩二[*1] Kohji Dohsaka　奥梓[*2] Azusa Oku　東中竜一郎[*3] Ryuichiro Higashinaka　南泰浩[*1] Yasuhiro Minami　前田英作[*1] Eisaku Maeda

[*1]日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation

[*2]大阪大学 大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University

[*3]日本電信電話株式会社 NTT サイバースペース研究所
NTT Cyber Space Laboratories, NTT Corporation

This paper investigates the evaluation functions of user engagement in thought-evoking spoken dialogues in which conversational agents promote user thinking and maintain user engagement in interactions under a wide range of dialogue topics. We collected the dialogue data by a WoZ experiment in which users engaged in quiz-style thought-evoking dialogues and signified the level of their engagement. Based on the user dialogue behavior extracted from the dialogue data, we grouped the users into three clusters by a partitional optimization clustering method and derived evaluation functions to predict user engagement in each cluster by stepwise multiple linear regression. We found that user clustering improved the performance of the evaluation functions compared to that derived from all users. We also found that the statistics of user behaviors over a series of dialogues and the distribution entropies of user behaviors over dialogue topics contribute to user engagement in different ways among the clusters.

1. はじめに

対話エージェントは，ユーザが言葉や身振りなどの日常的コミュニケーション手段を用いて容易にコンピュータとインタラクションを行うことを可能とするインタフェース技術である．人間・エージェント間のコミュニケーション手段としては様々なものがあるが，なかでも音声は日常で最もよく利用される手段の一つであり，これまで広く関心が寄せられてきた．

ユーザ・エージェント間の円滑で自然な音声対話を実現するためには対話の質の評価が重要な役割を果たす．PARADISE [Walker 00] は，対話データから抽出される評価尺度に基づいて対話の質のユーザ評定値を予測する評価関数を導出するための枠組みとして広く利用されている．PARADISE は，ステップワイズ変数選択方式の重回帰分析を適用することにより，評価尺度を説明変数としたときに，応答変数であるユーザ評定値を予測する評価関数を導出する．評価関数の性能は重回帰分析の決定係数 R^2 により測られる．

本稿では，ユーザ・エージェント間の思考喚起型音声対話において，収集した対話データに基づいてユーザ対話意欲を予測する評価関数を導出した結果について報告する．思考喚起型対話とは，エージェントがユーザの思考を喚起することで，ユーザの対話意欲を維持しようとする対話である．そうした対話の一例として，広い対話トピックの下でユーザ思考を喚起できるクイズ型音声対話 [Higashinaka 07, Minami 07] を取り上げる．思考喚起型対話は，できるだけ長くユーザ意欲を維持することが目的であり，できるだけ短い対話で効率的にタスクを遂行することを重視するタスク指向型対話とは観点が異なる．

多くの従来研究が PARADISE の枠組を様々な種類の対話に適用してきた [Foster 09, Litman 02, Möller 08, Rieser 10]. その中で，タスク指向型対話では 0.39 から 0.56 の決定係数 [Walker 00] や 0.71 の決定係数 [Litman 02] をもつ評価関数が報告されてきた．これに対し，効率的なタスク遂行を重視しない対話やタスクの成功・失敗が明瞭に区別できない対話では低い性能の評価関数が報告されている．例えば，ユーザ・ロボット間の協同タスク遂行対話では 0.20 の決定係数 [Foster 09]，情報推薦型対話 [Rieser 10] では 0.26 の決定係数の評価関数が報告されている．Möller 等は，タスク成功についてのユーザ主観評価を説明変数として使わない限り，重回帰分析に基づく評価法では，従来のタスク指向型対話で報告されてきた性能の評価関数を得ることは難しいことを示している [Möller 08].

この理由の一つとして，効率的タスク遂行が重視されない，あるいは，タスク成功の基準が明瞭でない対話においては，ユーザがタスクを効率的に遂行するという共通の目的に向かって行動する傾向が弱まり，ユーザ行動のばらつきが大きくなることが考えられる．このことがユーザ全体を説明する評価関数の性能を低下させている可能性がある．本稿で扱う思考喚起型音声対話においても同様の困難さが予測される．

この問題を解決するため，本稿ではユーザ対話行動の特徴に基づいてユーザを複数のクラスタに分割し，クラスタごとにユーザの対話意欲を予測する評価関数を導出する．さらに，評価関数導出の際の説明変数として，連続する複数の対話（対話系列）を通してのユーザ対話行動の統計量と対話トピック（クイズのカテゴリ）ごとのユーザ対話行動の分布エントロピーとに着目し，その効果を分析する．これは，思考喚起型対話では，複数の対話を通してできるだけ長くユーザ対話意欲を維持することが目的となることと，ユーザの関心と対話トピックの間の適合度がコミュニケーションの質に影響を与える可能性が高いことを重視したためである．

以下において，2 節で思考喚起型音声対話システムについて説明し，3 節で対話データ収集のために実施した Wizard-of-Oz(WoZ) 方式の対話実験について述べる．4 節で評価関数に

連絡先: 堂坂浩二, NTT コミュニケーション科学基礎研究所, 〒 619-0237 京都府相楽郡精華町光台 2-4, Email:dohsaka.kohji@lab.ntt.co.jp

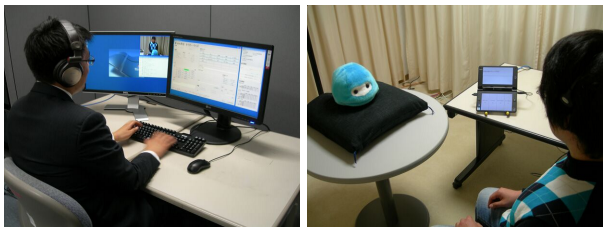

図 1: Wizard-of-Oz (WoZ) 対話実験風景: 左がエージェントを操作するオペレータ，右がエージェントと対話するユーザ．

ついて分析する．

2. 思考喚起型音声対話システム

対話データを収集するためにクイズ型音声対話システムを用いた [Higashinaka 07, Minami 07]．クイズ型対話は思考喚起型対話の一例であり，広い対話トピック（クイズのカテゴリ）の下でユーザの思考を喚起し，コミュニケーションを活性化することができる．対話エージェントは，ユーザに対して人名当てクイズを出題し，ヒントを順に提示していく．人名当てクイズのヒントは，容易には正解に辿り着かないように，難しいヒントから易しいヒントの順に自動的に並び替えられ，このことによりユーザの思考が喚起され，ユーザの対話意欲が向上することが分かっている [Higashinaka 07]．

システムは，音声認識によりユーザ発言を認識し，ユーザ発言の認識結果に基づいてエージェントの次発話を決定し，音声合成によりエージェント音声を出力する．エージェントは，物理的ロボットとして実現され，顔向きにより他の参加者の顔を見るという動作を行う．このようにシステムは自動的に対話を実行できるが，本実験においては，音声認識誤りの影響を回避するため，人間 (オペレータ) がエージェントの行動を操作するという WoZ 方式の対話実験を行った．図 1 に実験風景を示す．対話実験の詳細は 3 節で述べる．

図 2 に対話例を示す．エージェントは，ヒント提示 (1, 5)，ユーザ回答の評価 (3, 9, 11) を行うことができる．この他，ユーザの呼びかけやクイズの困難さや容易さの表出のための発言などを行うことができる．ユーザの発言は，回答発言 (2, 8, 10) と，回答以外の非回答発言の 2 つの種別に分類される．非回答発言としては，思考中であることの表出 (4)，クイズの困難さの表出 (6)，感想の表明 (7, 12) を行うことができる．この他，ユーザはクイズを途中でやめることもできる．非回答発言はさらに肯定，否定，中立の 3 つの種別に分類される．肯定発言にはクイズの容易さの表出や肯定的な感情をもつ感想の表明などがある (12)．否定的発言にはクイズの困難さの表出 (6)，否定的な感情をもつ感想の表明 (7)，クイズ放棄などがある．中立的な発言には思考中であることの表出 (4) とその他感情的に中立な感想の表明がある．

3. WoZ 対話実験

WoZ 方式の対話実験にはユーザとして 30 名 (男性:15 名，女性:15 名)，オペレータとして 2 名の日本語話者が参加した．

実験開始前に，クイズの問題となる有名な人物を芸能，スポーツ，文学・芸術，政治，歴史上の人物という 5 つのカテゴリから選んだ．カテゴリ間でクイズの難易度を揃えるため，日本語 Wikipedia [*1] 上のハイパーリンク構造に基づいて，Wikipedia

*1 http://ja.wikipedia.org/

1	エージェント:	なぞなぞスタート．最初のヒント．東京大学卒業だよ．(ヒント)
2	ユーザ:	吉田茂かな？ (回答)
3	エージェント:	全然違う．政治家じゃないよ．(回答評価)
4	ユーザ：	考え中です．(思考中 ／ 中立)
5	エージェント:	次のヒント出すよ．第 2 ヒント．小説家で，英文学の研究者だよ．(ヒント)
6	ユーザ:	うーん，難しい．(困難さ表出 ／ 否定)
7	ユーザ:	いらいらする．(感想 ／ 否定)
8	ユーザ:	村上春樹．(回答)
9	エージェント:	惜しい．小説家というのは合ってるよ．(回答評価)
10	ユーザ	夏目漱石．(回答)
11	エージェント:	あたり．すごい！よくわかったね．(回答評価)
12	ユーザ:	正解できてうれしい．(感想 ／ 肯定)

図 2: 対話例

中の人物の PageRank ™ スコアを計算し，人物をスコアの降順に並べた．次に，カテゴリ間で人物のスコアの分布がほぼ等しくなるように，カテゴリごとに 20 名の人物を選んだ．

実験においてユーザとオペレータは異なる部屋から対話に参加した．ユーザはエージェントと対面し，マイクロフォンを使って音声入力を行った．ユーザにはエージェントがオペレータにより操作されていることを知らさなかった．ユーザは，小休憩を挟みながら，2 時間程度対話を続けるように指示された．

各対話において，オペレータは 5 つのカテゴリからクイズの問題となる人物を選び，クイズを開始した．対話遂行の間，オペレータはヘッドフォンでユーザの音声を聞き，ユーザ発言の種別を分類し，その種別をキーボードでシステムに入力した．ユーザ発言の種別が回答の場合は，回答として発言された人物名も入力した．オペレータの入力情報とシステムが保持する対話文脈に基づいて，システムはエージェント発言の候補を出力した．オペレータはその候補からエージェント発言を一つ選び，発言はエージェントから音声で発声された．オペレータは，ユーザの対話意欲をできるだけ維持するように，クイズの問題となる人物の選択とエージェント発言の選択を行うように指示された．

ユーザは，1 クイズ対話が終了するごとに，その時点での対話意欲の程度を 4 段階 (1〜4) で評定し，タッチパネルで入力した．4 が対話意欲が最も高く，1 が最も低い．対話意欲のユーザ評定値はオペレータには知らされなかった．

各ユーザは平均 38.8 回のクイズ対話を行い，合計 1,163 個の対話が収集された．1 対話の長さは平均 2.92 分であった．対話データ全体で 43,021 個の発言があった．そのうち，エージェントの発言は 27,439 個，ユーザの発言は 15,582 個であった．ユーザ対話意欲は平均 3.07，標準偏差 0.84 であった．

4. ユーザ対話意欲の評価関数の導出

4.1 評価尺度

対話データから抽出される評価尺度に基づいて，各対話終了時点のユーザ対話意欲を予測する評価関数を導出した．まず，評価尺度について説明する．評価尺度は対話効率，対話の質，タスク成功の観点から分類される [Walker 00]．評価尺度として，現在の対話に限ったユーザ行動の統計量と，それまでに実行された複数の対話から成る対話系列を通したユーザ行動の統計量の両方を検討し比較した．ここで，対話系列を通したユー

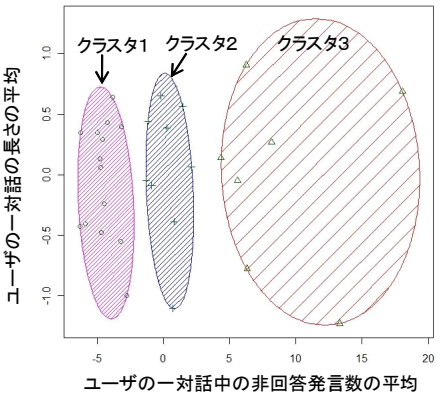

図 3: ユーザクラスタリングの結果．見やすさのため，軸のスケールは調整されている．

ザ行動は，対話系列中のユーザの平均的な振る舞いと，現在の対話のユーザ行動が一つ前の対話までのユーザの平均的振る舞いからどう変化したかを表わす差によってモデル化した．

第一に，対話の効率性に関する尺度として次を検討した．

(E1) 現在の目標達成対話の長さと目標不達成対話の長さ

(E2) それまでの対話系列における目標達成対話の長さの平均値と目標不達成対話の長さの平均値

(E3) 現在の対話の長さから，1 つ前の対話までの対話系列における対話の長さの平均値を引いた差

対話の長さは対話の所要時間 (分) によって測った．クイズが正解して終了した対話を目標達成対話，それ以外の対話を目標不達成対話と定義する．(E1) は，現在の対話が目標達成／不達成対話ならばその対話の長さであり，さもなければ 0 となる尺度である．(E1) が現在の対話の統計量であり，(E2) と (E3) が対話系列を通した統計量である．

第二に，対話の質に関する尺度として次を検討する．

(Q1) ユーザの各種発言に関して，現在の対話における発言数

(Q2) ユーザの各種発言に関して，それまでの対話系列を通しての一対話あたりの発言数の平均値

(Q3) (Q1) から一つ前の対話までに実行された対話における発言数の平均値を引いた差

(Q4) 対話トピック（クイズのカテゴリ）ごとの目標達成対話の分布エントロピー

これらの尺度を取り上げたのは，思考喚起型対話では，ユーザ発言数や，ユーザ関心と対話トピック分布の間の適合がコミュニケーション活性の程度と関係すると考えられるからである．ここで，ユーザの各種発言とは，クイズ型対話における回答発言，非回答発言，肯定発言，否定発言，中立発言のことを言う．発言数は対話の時間長さで正規化した．

第三に，タスク成功に関わる統計量として次を検討した．

(S1) 現在の対話が目標達成対話ならば 1，さもなければ 0 となる変数

(S2) 対話系列における目標達成対話の割合

(S3) (S1) から一つ前の対話までの対話系列における目標達成対話の割合を引いた差

対話	全ユーザ	クラスタ 1	クラスタ 2	クラスタ 3
全て	0.11	0.32	0.33	0.62
	(0.092)	(0.18)	(0.23)	(0.61)
前半	0.17	**0.41**	0.23	**0.66**
	(0.14)	(0.27)	(0.007)	(0.62)
後半	0.073	**0.35**	**0.53**	**0.66**
	(0.035)	(0.22)	(0.16)	(0.53)

表 1: 評価関数の決定係数 R^2．各セルで上段は対話系列を通した統計値に基づく結果，下段の () 内は現在の対話の統計値に基づく結果．

4.2 評価関数

分割最適化クラスタリング手法の一つである PAM (Partitioning Around Medoids) 法を適用し，対話データから抽出したユーザ対話行動の統計量に基づいてユーザのクラスタリングを行った．一連の対話の冒頭でユーザのクラスタリングを行うことを想定して，クラスタリングには各ユーザの最初の 5 対話から抽出された統計量を用いた．クラスタ数と統計量は，生成されるクラスタのシルエット係数が 0.5 以上となり，各クラスタに属するユーザ数が 3 を越えるものを選んだ．結果として，図 3 に示すように，ユーザは，ユーザの 1 対話中の非回答発言と 1 対話の長さの平均に基づいて，3 つのクラスタに分割された: クラスタ 1 (14 ユーザ), クラスタ 2 (9 ユーザ), クラスタ 3 (7 ユーザ).

各ユーザの 6 番目以降の対話を対象として，ステップワイズ変数選択による重回帰分析を適用し，全ユーザとユーザクラスタごとにユーザ対話意欲を予測する評価関数を導出した．長時間のインタラクションの中でユーザの振る舞いが変化する場合を考慮して，各ユーザの対話を前半と後半に分割し，全対話・前半の対話・後半の対話のそれぞれについて評価関数を導出した．前半の対話は 6 番目から 22 番目の対話とし，後半の対話は 22 番目以降の対話とした．

対話系列を通してのユーザ対話行動の統計量 (E1,Q1,T1) と現在の対話のユーザ対話行動の統計量 (E2,E3,Q2,Q3,Q4,S2,S3) の効果を比較した．

表 1 は，評価関数の決定係数 R^2 を示している．この決定係数は 10-fold cross validation 法による平均値である．決定係数は評価関数の性能を表す．表の各セルにおいて，上段は対話系列を通した統計値に基づく結果，下段のは現在の対話の統計値に基づく結果を示す．

明らかにユーザクラスタリングにより評価関数の性能が改善されていることが分かる．予測された通り，タスク遂行の効率性が重視されないクイズ型の思考喚起型対話では，ユーザのクラスタリングを行わず全ユーザの対話データから評価関数を導出した場合，評価関数の性能は低い．このことは，対話意欲の観点から見たとき，ユーザの対話行動の傾向は多岐に渡るが，ユーザを適切にグループ化することにより対話意欲の予測が容易になることを示している．

さらに，現在の対話に限った統計量に比べて，対話系列を通した統計量は，ユーザ対話意欲の予測により貢献している．また，対話の前後半への分割は，クラスタ 2 の前半の対話を除いて，評価関数の性能を改善する．この結果は，思考喚起型対話の目的がユーザ対話意欲をできるだけ長く維持することにあることを考えると，もっともらしい結果であり，ユーザ対話意欲が，現在の対話だけでなく，過去の対話系列を通した対話行動の履歴に大きく依存することを示すものと考えられる．

表 2 は，対話系列を通した統計量に基づく各評価関

対話	クラスタ	評価関数	R^2
前半	クラスタ 1	0.33 * $\mathcal{L}$(SuccessfulDialog) - 0.42 * $\mathcal{L}$(FailedDialog) + 0.35 * $\mathcal{N}$(Positive) + 0.21 * $\mathcal{N}$(Neutral) + 0.16 * H_G(SuccessfulDialog)	0.41
前半	クラスタ 2	0.17 * $\mathcal{L}$(SuccessfulDialog) + 0.63 * $\mathcal{N}$(Positive) + 0.57 * $\mathcal{N}$(Negative) - 0.19 * H_G(SuccessfulDialog)	0.23
前半	クラスタ 3	0.18 * Δ_N(Answer) + 0.88 * $\mathcal{N}$(Neutral) - 0.23 * H_G(SuccessfulDialog)	0.66
後半	クラスタ 1	0.46 * $\mathcal{L}$(SuccessfulDialog) - 0.20 * $\mathcal{L}$(FailedDialog) - 0.33 * $\mathcal{N}$(Answer) + 0.27 * $\mathcal{R}$(SuccessfulDialog) + 0.30 * H_G(SuccessfulDialog)	0.35
後半	クラスタ 2	0.70 * $\mathcal{N}$(Answer) + 0.16 * $\mathcal{N}$(Positive) + 0.89 * $\mathcal{N}$(Negative) - 0.25 * $\mathcal{N}$(Neutral) - 0.22 * $\mathcal{R}$(SuccessfulDialog)	0.53
後半	クラスタ 3	- 0.33 * $\mathcal{L}$(SuccessfulDialog) - 0.17 * Δ_L(Dialog) - 0.25 * $\mathcal{N}$(Positive) + 0.90 * $\mathcal{N}$(Neutral)	0.66

表 2: 評価関数における説明変数の標準化偏回帰係数

数を説明変数の標準化偏回帰係数とともに示している．$\mathcal{L}$(SuccessfulDialog) と $\mathcal{L}$(FailedDialog) は，それぞれ目標達成対話と目標非達成対話の長さの平均値 (E2) を表わす．Δ_L(Dialog) は差 (E3) を表わす．$\mathcal{N}$(Answer), $\mathcal{N}$(Positive), $\mathcal{N}$(Negative), $\mathcal{N}$(Neutral) は，それぞれ回答のための発言，肯定発言，否定発言，中立発言に関して，単位時間あたりの発言数の対話系列を通した平均 (Q2) を表わす．Δ_N(.) は差 (Q3) を表わす．H_G(SuccessfulDialog) はエントロピー (Q4) を表わす．$\mathcal{R}$(SuccessfulDialog) は目標達成対話の割合 (S2) を表わす．差 (S3) はステップワイズ変数選択による重回帰分析を適用する過程において消去された．

各説明変数の貢献の大きさと正負の向きはクラスタごとに変化しており，対話意欲に依存するユーザ行動の傾向はクラスタにより異なることが分かる．第一に，クラスタ 1 において，目標達成対話の長さの平均はユーザ対話意欲に対して正の向きに働くが，クラスタ 3 の後半の対話においては負の向きに働いている．第二に，クラスタ 3 では，中立の非回答発言数の平均がユーザ対話意欲に対して正の向きに大きく働いていることが分かる．図 3 で示したように，クラスタ 3 のユーザは非回答発言を多く行う傾向があるが，その中でも中立の非回答発言がユーザ対話意欲と密接に関係していることが示されている．第三に，対話トピックごとの目標達成対話の分布エントロピー (Q4) は，ユーザ対話意欲の予測に貢献する場合があるが，その働きの向きはクラスタによって異なる．すなわち，エントロピー (Q4) はクラスタ 1 においてはユーザ対話意欲に対して正の向きに働くが，クラスタ 2 の前半においてはユーザ対話意欲に対して負の向きに働いている．このことは，広い対話トピックを好むユーザと狭い対話トピックを好むユーザが存在する可能性を示唆している．

5. おわりに

本稿では，思考喚起型対話の一例として，広い対話トピックの下でエージェントがユーザの思考を喚起することにより，ユーザ対話意欲を維持することができるクイズ型音声対話に着目し，そうしたクイズ型音声対話においてユーザ対話意欲を予測する評価関数について分析した．WoZ 対話実験により収集した対話データを用い，ユーザの 1 対話あたりの非回答発言数の平均と対話の長さの平均に基づいて，ユーザをクラスタに分割し，クラスタごとにユーザ対話意欲の評価関数を導出した．その結果，ユーザのクラスタリングにより評価関数の性能が向上することが分かった．加えて，対話系列を通したユーザ対話行動の統計量と，対話トピック（クイズのカテゴリ）ごとの目標達成対話の分布エントロピーが，ユーザクラスタごとに様々にユーザ対話意欲の予測に貢献することを示した．今後の課題としては，ユーザの言語行動に加えて，非言語行動も利用してユーザ対話意欲を予測することがある．

謝辞：本研究の一部は，科研費（新学術領域）「人とロボットの共生による協創社会の創成」における計画研究「ロボットのコミュニケーション戦略の生成」(21118004) の助成を受けたものである．

参考文献

[Foster 09] Foster, M. E., Giuliani, M., and Knoll, A.: Comparing objective and subjective measures of usability in a human-robot dialogue system, in *Proc.ACL/AFNLP 2009*, pp. 879–887 (2009)

[Higashinaka 07] Higashinaka, R., Dohsaka, K., Amano, S., and Isozaki, H.: Effects of quiz-style information presentation on user understanding, in *Proc. Interspeech 2007*, pp. 2725–2728 (2007)

[Litman 02] Litman, D. J. and Pan, S.: Designing and Evaluating an Adaptive Spoken Dialogue System, *User Modeling and User-Adapted Interaction*, Vol. 12, No. 2-3, pp. 111–137 (2002)

[Minami 07] Minami, Y., Sawaki, M., Dohsaka, K., Higashinaka, R., Ishizuka, K., Isozaki, H., Matsubayashi, T., Miyoshi, M., Nakamura, A., Oba, T., Sawada, H., Yamada, T., and Maeda, E.: The World of Mushrooms: human-computer interaction prototype systems for Ambient Intelligence, in *Proc. ICMI 2007*, pp. 366–373 (2007)

[Möller 08] Möller, S., Engelbrecht, K.-P., and Schleicher, R.: Predicting the quality and usability of spoken dialogue services, *Speech Communication*, Vol. 50, pp. 730–744 (2008)

[Rieser 10] Rieser, V., Lemon, O., and Liu, X.: Optimising information presentation for spoken dialogue systems, in *Proc. ACL 2010*, pp. 1009–1018 (2010)

[Walker 00] Walker, M., Kamm, C., and Litman, D.: Towards developing general models of usability with PARADISE, *Natural Language Engeneering*, Vol. 6, pp. 363–377 (2000)