1ZB-3

ロボットへの話しかけやすさのオンライン予測へ向けた検討

杉山 貴昭 † 駒谷 和範 † 佐藤 理史 †
† 名古屋大学大学院 工学研究科 電子情報システム専攻

1 はじめに

人間同士の対話には，対話者同士が無意識のうちに守っているルールが存在する．例として，人間は相手の状態を考慮して話しかけることや，相手の方向を向いて発話することが挙げられる．このようなルールを本研究では社会的規範と呼ぶ．我々は，人間に類似したロボットとユーザとの対話でも，ユーザは社会的規範を守りながら，ロボットと対話すると考える [1].

これまで我々は社会的規範の一部として，ロボットの一連の発話や挙動に対して，ユーザが話しかけられると感じるタイミングを，ロボットが予測するモデルを構築した [2]．話しかけやすさを予測する枠組みを図 1 に示す．入力は，任意の時点でロボット自身から得られる情報であり，例えば，ロボットの姿勢や動作，発話中か否かなどである．これらを用いてロジスティック回帰を行い，話しかけやすい，話しかけにくいの 2 値を出力する．本モデルを実際に人とロボットとの音声対話で利用すれば，雑音等に対して頑健な対話を実現できる．例えば，ロボットの発話中に雑音が発生した際に，ロボットがその時点での話しかけられやすさを考慮できれば，従来の入力音判別手法 [3] と組み合わせることで，この雑音をより高精度に棄却できる．

本稿では，このモデルを実際の音声対話で使用可能とするために，これをオンラインで予測する際の課題について述べる．ここでの課題は次の 2 点である．

1. 対話相手に応じた予測
2. 実際に発話する場合との違いの検証

まず，これまでのオフラインでの実験では，全てのユーザに対して平均的な性能で予測可能なモデルの構築を目指していた [2]．これでは不十分な理由は，人間はよく話しかける人やあまり話しかけない人のような属性（以降，ユーザの属性）をもっており，話しかけやすさの感じ方はユーザによって異なるからである．実際にユーザと音声対話を行う場合，このモデルが個々のユーザの属性に適応できる必要がある．次に，これまでのデータ収集は，ユーザが話しかけやすいと感じた時にマウスをクリックするという方法であった．一方，実際の音声対話では，ユーザは話しかけやすいと感じたタイミングで発話を行う．そのため，前者の方法で収集したデータが，実際にユーザが発話した場合に得られるデータと異なる可能性があった．

2 対話相手に応じた予測

対話相手に応じて話しかけやすさを予測するために，ユーザの属性に合わせて複数個の学習データを作成する．これらの出力の重み付き和を，対話中に推定したユーザの属性を利用して，2 値に判別する．

ユーザの属性に適応して話しかけやすさを予測するための枠組みを図 2 に示す．まず，話しかけやすい，話しかけにくいとしたサンプル数の異なる複数の学習データ

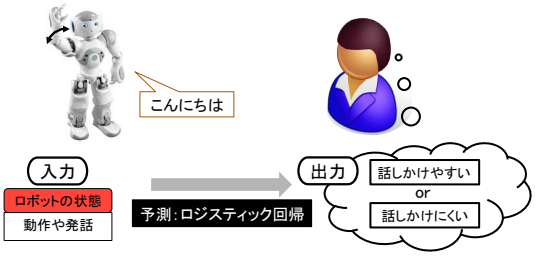

図 1: 話しかけやすさを予測する枠組み

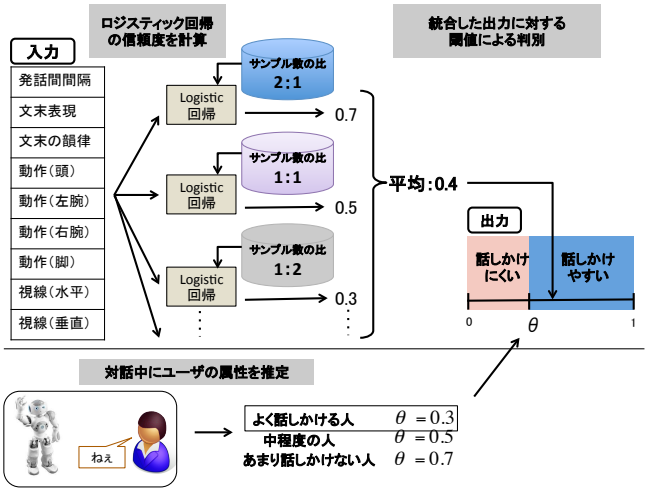

図 2: ユーザの属性への適応のための枠組み

を用意する．これにより，例えば，話しかけやすいとしたサンプル数が多い学習データで作成したモデルは，より多く話しかけやすいと出力する．このように作成したモデルの出力は，サンプル数の比によって変動する．次に，これらの複数の出力の平均をとる．最後に，得られた値をある閾値によって，話しかけやすさを出力する．この閾値はユーザの属性によって変動させ，例えば，対話中に得られる情報（例えば，一定時間内での発話回数）から自動で推定する．

学習データには，論文 [4] で収集したデータを用いる．このデータは，25 名の一般ユーザにロボットの一連の挙動を通して見せ，GUI を用いて話しかけやすさを付与させたものである．ロボットの一連の挙動の長さは 259.3 秒であり，これにより収集したデータを 0.1 秒毎にサンプリングし，2593 個のデータとした．

各学習データ間で，2 つのラベルのサンプル数の比を変えるため，学習データの正解ラベルは，N 人以上が話しかけやすいとした区間を「話しかけやすい」，それ以外を「話しかけにくい」とした．N の値とその時の 2 つのラベルのサンプル数の関係を図 3 に示す．縦軸はサンプル数，横軸は N の値である．図のように，N が小さいと話しかけやすいのサンプル数が多く，N が大きいと話しかけにくいのサンプル数が少ないことがわかる．そこ

Online Prediction of How Likely User is to Talk to Humanoid Robot: *Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato (Nagoya Univ.)*

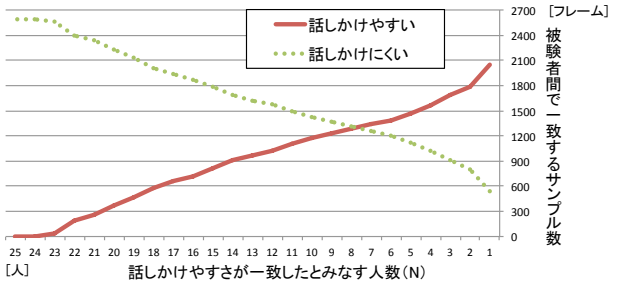

図 3: 話しかけやすさが一致した人数とその時のサンプル数の関係

で，話しかけやすいのサンプルがなかった $N = 24, 25$ の場合を除く，$1 \leq N \leq 23$ までの N が奇数番目の 12 個の学習データを利用した．これらの学習データを用いて，ロジスティック回帰の重みを学習し，12 個のモデルとした．

3 実際に発話する場合との違いの検証

本研究では，新たに発話によるデータ収集を行う．これにより，以前のデータ収集の方法で，実際にユーザが発話した場合と同様のデータが得られることを確かめる．

データ収集では，6 名の被験者に発話とマウスによる両方の方法で話しかけやすさを付与させた．同一被験者に両方の方法で実験させたのは，個人差による影響を防ぐためである．3 名の被験者には，まずマウスをクリックする方法，次に発話する方法という順で行い，残り 3 名は逆の手順で行った．この時に使用した挙動や教示と，マウスをクリックする方法は，論文 [4] と同一である．

発話によるデータ収集として，被験者にロボットの挙動を見せ，話しかけられると感じたタイミングで「ねぇ」と発話させた．被験者の音声は，ロボットの動作音や雑音の混入を防ぐため，ヘッドセットマイクを用いて収集し，Julius[‡]付属の adintool で録音した．収録した音に対して，Julius により音声認識を行い，ユーザの発話開始時刻を取得した．

4 評価実験

4.1 対話相手に応じた予測に対する性能評価

ユーザの属性に適応できれば，話しかけやすさの予測性能が向上することを示す．具体的には，事後的に閾値を変更し，ユーザ毎の予測性能を確かめる．我々は今後オンラインで予測することを想定しているため，本来ならば，自動でユーザの属性をロボットが推定できる必要がある．将来的には，ユーザの属性は，当該ユーザの一定時間内での発話頻度など，対話中の情報から推定できると考えている．

テストデータとして，学習とは異なる挙動を用いて収集したデータを用いる．この挙動の長さは 150.0 秒であり，学習データと同様の方法で 1500 個のデータとした．データ収集では，25 名の被験者に対して 3 回ずつデータを収集しており，これら全てのデータをテストデータとする．評価指標として，「話しかけやすい」「話しかけにくい」の正解ラベルと，ロジスティック回帰の出力が一致した数から，MacroF1[§]を計算する．

[‡]http://julius.sourceforge.jp/

[§]「話しかけやすい」の F 値と「話しかけにくい」の F 値の平均値．F 値は再現率と適合率の調和平均．

表 1: 複数のモデルを統合するか否かの性能比較

	論文 [4]	提案手法	
統合	なし	あり	
閾値	最適化	固定 (0.3)	最適化
MacroF1	75.1	71.0	72.4

表 2: クリックと発話の一致度合

被験者	1	2	3	4	5	6	平均
$N(u)$	48	17	35	18	21	19	
$N(c)$	43	13	24	15	18	15	
$N(c)/N(u)$	89.6	76.5	68.6	83.3	85.7	78.9	80.4

表 1 に，複数のモデルを統合するか否かの性能比較を示す．比較対象として，21 名以上の共通区間を学習データとして利用した，論文 [4] の手法を用いた．表の結果は，被験者 25 名が各 3 回ずつ試行した計 75 回のデータに対して予測した時の MacroF1 を示している．閾値最適化は，試行毎に 0.1 刻みで 0.1 から 0.9 まで変化させ，最も MacroF1 が高い値を選択した．なお，同一条件にするため，比較対象でもロジスティック回帰の閾値を試行毎に最適化した．また，閾値適応なしは，事後的に MacroF1 の平均が最も高くなる閾値（0.3）に固定した場合である．表より，複数のモデルを統合した場合の性能は，固定した場合よりは高いものの，論文 [4] の手法で閾値を最適化させたほうが高いことがわかった．提案手法の性能が低い理由の 1 つに，複数の学習データで作成したモデルの各出力が，サンプル数の比によって上手く変動していなかったことが考えられる．

4.2 発話する場合との違いの検証に対する調査結果

3 章で得られたデータを比較し，マウスをクリックする方法でも同様のデータが得られることを確認する．収集したデータは，6 名の被験者が 2 つの方法で 2593 フレームに対して話しかけやすさを付与したデータである．

評価方法として，マウスをクリックしている区間内に，ユーザの発話開始時刻がどのくらい含まれているかを用いる．同一ユーザから 2 つの方法で収集したデータの一致度合を，表 2 示す．ここで，発話回数を $N(u)$，クリック中の発話数を $N(c)$，クリック中に発話した割合を $N(c)/N(u)$ とする．表より，クリック中に発話した割合の全被験者に関する平均は 80.4%であった．これにより，マウスをクリックする方法でも，発話する方法と同様のデータが得られることを示した．つまり，本研究で作成した学習データは，オンラインで話しかけやすさを予測する際でも利用可能であることが分かった．

参考文献

[1] B. Reeves and C. Nass. The media equation: How people treat computers, televisions, and new media as real people and places. *Cambridge University Press*, 1996.

[2] 杉山貴昭, 駒谷和範, 佐藤理史. ヒューマノイドロボットが話しかけやすさを予測するモデルの構築. 人工知能学会論文誌, Vol. 28, No. 3, pp. 255–266, 2013.

[3] A. Lee, K. Nakamura, R. Nisimura, H. Saruwatari, and K. Shikano. Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs. *Proc. Interspeech*, pp. 173–176, 2004.

[4] 杉山貴昭, 駒谷和範, 佐藤理史. ロボットへの話しかけやすさモデルの評価と個人差や教示による変動への対応. 人工知能学会論文誌, Vol. 29, No. 1, pp. 32–40, 2014.