観測した周囲の状況を曖昧な発話に統合した対話ロボットによる気の利いた行動選択

田中翔平 [1,2] 山崎康之介 [1,2] 湯口彰重 [2,1] 河野誠也 [2] 中村哲 [1] 吉野幸一郎 [2,1]
[1] 奈良先端科学技術大学院大学 [2] 理化学研究所 ガーディアンロボットプロジェクト
{tanaka.shohei.tj7, yamasaki.konosuke.yi5, s-nakamura}@is.naist.jp
{akishige.yuguchi, seiya.kawano, koichiro.yoshino}@riken.jp

概要

人と協働する対話ロボットは，ユーザの要求が曖昧な場合でもユーザが必要とする行動を取ることが期待される．言い換えれば，対話ロボットはユーザ発話の内容のみでなく，ユーザの周囲の状況を正確に理解して気の利いた行動を選択する必要がある．本研究では，このように周囲の状況から得られるマルチモーダル情報を活用するロボットの行動選択モデルを構築した．また，モデルに入力するユーザ発話や周囲の状況を人手で書き起こした場合と自動認識した場合を比較した．実験結果より，気の利いた行動をとるために必要な周囲の状況を選択的に与えることで行動の選択精度が向上することがわかった．また，周囲の状況に関して精度に限りがある自動認識結果を用いた場合でも，単純な事前学習モデルで抽出した特徴量を用いる場合よりは高精度で行動を選択できることが明らかになった．

1 はじめに

対話ロボット・システムが人と協働することを想定したこれまでの研究の多くは，ユーザの要求がロボットに対して明示される，あるいはシステムの問い返しによって要求が明確化されることを仮定していた [1, 2]．しかし実際には，ユーザ自身が持つ要求が曖昧で，ユーザから明示的な要求を示すことができない場合も多い [3, 4]．“曖昧” とはユーザが何らかの潜在的な要求を持っているにも関わらず，その要求の条件を明確に言語化できない状況にあることを意味する [5]．こうした曖昧な要求に対して，気心の知れた人間同士であれば気を利かせて相手が必要としそうな補助を行動として起こすことができる．例えば，人が起きたタイミングで水を持っていく，ため息をついたときに「どうしましたか？」などと聞くような行動を取ることができる．このように，人間と生活環境を共にする対話ロボットは，単に相手からの要求に応じて動作するだけではなく，ユーザ発話に紐づいた状況に応じて能動的に振る舞いを決定することが求められる．

こうした気の利いた行動をとることができるシステムを実現するため，これまでにユーザの曖昧な要求とユーザの周囲の状況を表す画像，およびロボットの気の利いた行動で構成されるコーパスを収集してきた [6]．このコーパスは，ユーザ発話のテキストやロボットの一人称視点を想定した画像というマルチモーダルな情報を統合的に活用し，ロボットがそのユーザの状況にあった行動 (気の利いた行動) を選択することを想定している．このデータで，入力をユーザの発話と周囲の状況，出力を気の利いた行動として教師ありで分類するベースラインモデルを構築した．このとき，特に状況についてシステムが正しく認識をできるよう，様々な説明的な特徴量のアノテーションを入力として利用した．これは例えば，ユーザ発話時点でのコーヒーテーブル上の物体，ユーザの把持物体，ユーザの姿勢などが含まれる．こうした特徴量をベースラインモデルの入力として利用した場合，大規模事前学習モデルを用いて抽出した特徴量のみを与える場合よりも気の利いた行動選択の精度が大きく向上することがわかった．

しかし，実際にロボットでこうした特徴量を利用する場合，状況に対する説明的な特徴量をいかに自動認識して利用できるかという問題が生じる．そこで本研究ではこれらの特徴量を自動認識した場合にどの程度の精度で気の利いた行動を選択できるか検証した．この検証には，一般に広く用いることができる音声認識，物体認識，姿勢推定などのモデルを利用した．実験結果より，説明的な特徴量を自動認識した場合，アノテーションで与えた特徴量を使う

図 1　気の利いた行動の例: ロボットはバナナを持ってきている．

表 1　家庭内ロボットの行動カテゴリリスト

バナナを持ってくる，充電ケーブルを持ってくる，コップを持ってくる，ケチャップを持ってくる，宅配便を持ってくる，ペットボトルを持ってくる，リモコンを持ってくる，スマホを持ってくる，お菓子を持ってくる，ティッシュ箱を持ってくる，充電ケーブルを片付ける，コップを片付ける，ケチャップを片付ける，ミニカーを片付ける，ペットボトルを片付ける，リモコンを片付ける，スマホを片付ける，お菓子を片付ける，ティッシュ箱を片付ける，ゴミをゴミ箱に捨てる，缶切りを持ってくる，クッキングシートを持ってくる，グラスを持ってくる，おろし器を持ってくる，キッチンペーパーを持ってくる，レモンを持ってくる，オリーブオイルを持ってくる，じゃがいもを持ってくる，サランラップを持ってくる，水筒を持ってくる，缶切りを棚にしまう，クッキングシートを棚にしまう，グラスを棚にしまう，おろし器を棚にしまう，キッチンペーパーを棚にしまう，ペットボトルを冷蔵庫にしまう，サランラップを棚にしまう，タッパーをレンジに入れる，タッパーを冷蔵庫にしまう，水筒を棚にしまう

場合よりも行動選択の精度が大幅に低下するものの，事前学習モデルを用いて抽出した特徴量のみを用いる場合と比較して高い精度が実現されることがわかった．

2　タスク設定とコーパス

本研究で取り組む課題は，一般的なリビングやキッチンにおいてロボットがユーザの家事を手伝うという状況を想定したものである．ユーザは要求が曖昧な発話や独話を行い，ロボットはユーザ発話とユーザ発話が行われた状況を見ながら気の利いた行動をとる状況を想定する．図 1 にユーザとロボットのインタラクションの例を示す．ここでユーザの「朝食が食べたい」という発話は，必ずしも特定の機能に対する要求として言語化されているわけではない．これに対して，ロボットはユーザ発話と机の上に“バナナ”があるなどの状況を勘案しつつ「バナナを持ってくる」という気の利いた行動を選択し，実際にバナナをユーザのもとに持ってくる．

こうした状況を想定して，ユーザの曖昧な発話，その発話が行われた状況を表すロボットの一人称視点を想定した画像，それに対応したロボットの気の利いた行動の三つ組で定義されたコーパスを収集した [6]．このコーパスにおいて，ユーザの曖昧な発話が入力されたとき，ロボットはその発話状況と発話内容の情報を活用し，あらかじめ定義された行動カテゴリの中から気が利いているとみなせるような行動カテゴリを出力する．あらかじめ定義された全 40 種類の行動カテゴリのリストを表 1 に示す．具体的には，どのような状況でロボットがこの行動をとったら気が利いていると思うかをクラウドワーカーに尋ね，ユーザの先行発話および室内の状況を入力してもらうことで状況-行動ペアを収集した [4]．

また，室内で家事を補助するロボットがユーザの曖昧な要求を受けたときに，発話内容のみからでは適切な気の利いた行動を選択することが難しい場合が存在する．例えば「あれ？もう一個足りない」という発話はユーザがグラスとお酒を持っている状況で，もう一つグラスが欲しいという状況における発話である．このとき発話内容のみから適切な行動である「グラスを持ってくる」を選択することは困難であるが，「ユーザがグラスを手に持っている」という画像中の情報を参照することで紐づけが可能となる．

また，ある環境でのロボットの一人称視点を想定したデータを用いることを想定したとき，大量のデータを用意することは難しい．そこで，画像から得られる様々な抽象化レベルの情報を効率よく利用するため，状況を説明するようなラベル (説明的な特徴量) の付与を人手で行った．この際，動画から最後のフレームの画像を代表画像としてクリップし，代表画像における物体や人物姿勢などのラベルを付与した．

収集されたユーザ発話および発話に紐付けられた画像，画像から得られる説明的な特徴量の例を図 2 に示す．*Uttr* はユーザ発話を意味し，*action* は対応する気の利いたロボットの行動を意味する．*Viewpoint* は画像を撮影したカメラの視点番号を意味し，計 3 種類である．*Position* はソファやキッチンなど，ユーザが室内のどこにいるかを表す特徴量である．*Pose* は座っている，立っているといったユーザの姿勢を表す特徴量である．*Has* はユーザが持っている物体を表す特徴量である．*Coffee table* はコーヒーテーブル上に置かれた物体を表す特徴量である．*Dining table* はダイニングテーブル上に置か

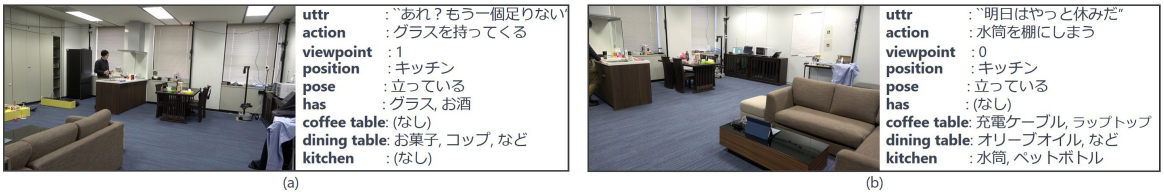

図 2　コーパスに含まれる対話例

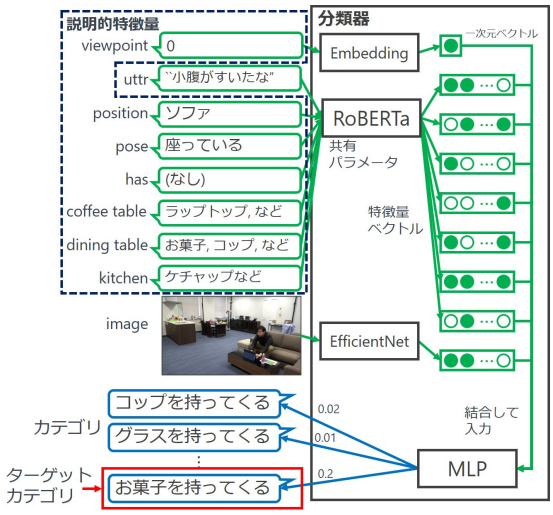

図 3　ベースライン分類器における特徴量の入力

れた物体を表す特徴量である．*Kitchen* はキッチンに置かれた物体を表す特徴量である．これらの特徴量は人手で付与されているが，画像認識などで自動で抽出することを指向したデザインになっている．これまでに提示したデータの収集には多大なコストが掛かるため，本コーパスでは 400 件の事例が収録されている．

3　マルチモーダル情報を活用する気の利いた行動の分類器

収集したデータを用い，曖昧なユーザ発話と状況から気の利いたロボット行動を推定するベースラインモデルを構築する．図 3 にベースライン分類器の概要を示す．分類器は入力された状況に対して，その状況で気が利いているとみなすことができるロボットの正解行動を 40 クラスから選択する．各特徴量の具体的な処理及び正解カテゴリの予測については次の通りである．まずユーザ発話 (*uttr*) は事前学習モデルである RoBERTa [7] に入力し，RoBERTa が出力した [CLS] ベクトルを特徴量ベクトルとする．また画像から抽出された特徴量である *position, has, coffee table, dining table, kitchen* はテキストであるため，これも [SEP] トークンで結合してそれぞれ RoBERTa へ入力し [CLS] ベクトルを特徴量ベクトルとする．*Viewpoint* については各 *viewpoint* に対応する一次元ベクトルを embedding 層より取得する．これらの特徴量は画像から得られる説明的な特徴量 (description) である．これ以外にも，一般に用いられる画像の事前学習モデルを利用した特徴量として，画像 (*image*) を EfficientNet-B0 [8] に入力してベクトル化する．出力層においては，これらの手続きによって得られた特徴量ベクトルをすべて結合し，1 層の Multi Layer Perceptron (MLP) へと入力して各カテゴリに対応する確率値を算出する．

4　実験

行動選択のために状況に対する説明的な特徴量を用いる場合，実際にその特徴量が認識可能かという問題が生じる．そこで本節ではこれらの特徴量が人手で与えられた場合と自動認識の結果として与えられた場合の性能変化を調査する．ユーザ発話の音声認識は Google Speech-to-Text API[1] を用いた．また説明的な特徴量の認識には EfficientNet+MLP を用いた．説明的な特徴量の全クラスは既知のクラスであると設定し学習を行った．特徴量認識の性能を表 2 に示す．*Viewpoint* はロボット自身の現在位置から一意に定まる値であるため推測していない．また *has* はそもそも含まれているインタラクションの数が少数であり学習が不可能であったため表 2 には含まれていない．音声認識の評価には word error rate (WER), match error rate (MER), word information lost (WIL) [9] を用いた．音声認識の精度は中程度であるが，これはロボットがユーザから 2m ほど離れた距離で音声認識を行う状況を想定した結果であることに注意されたい．説明的な特徴量についてはどれも高い精度で認識できており，特に *position, pose* については 100%に近い精度で認識できている．これはユーザの位置，ポーズはそれぞれ 3 種類と非常に限られているからだと考えられる．図 5 に *coffee table, dining table, kitchen* に含まれる各オブジェクトについて，データセット中における出現回数と認識されなかった割合 (認識失敗率) の散布

1）　https://cloud.google.com/speech-to-text

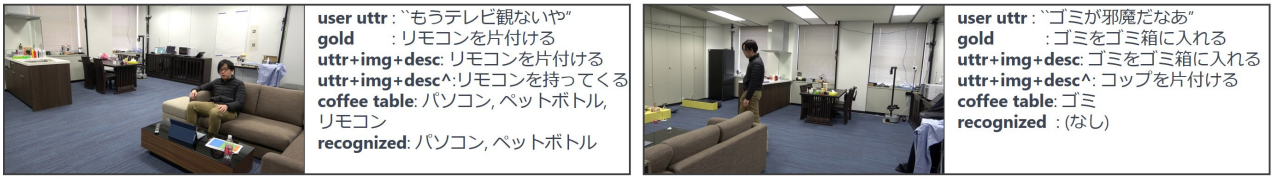

図 4 気の利いた行動の選択に重要なオブジェクトが認識されなかったインタラクションの例

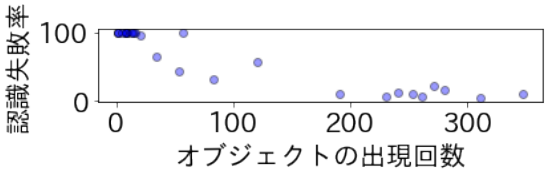

図 5 オブジェクトの出現回数と認識失敗率の散布図．相関係数は –0.93.

図を示す．相関係数は –0.93 であり，オブジェクトの出現回数と認識されなかった割合には強い負の相関があることがわかる．

これらの自動認識した特徴量を用いるモデルの性能を表 3 に示す．評価指標として，Accuracy (Acc.), Recall@5 (R@5), Mean Reciprocal Rank (MRR) [4] を用いた．*Uttr* は図 3 におけるユーザ発話のみをモデルへの入力とした場合を，*uttr+img* は発話に加え EfficientNet で作成した画像特徴を入力とした場合を，*uttr+img+desc* は *uttr+img* に加え画像から得られる説明的な特徴量 (description) を人手で与えた場合を意味する．“^ (hat)” はユーザ発話または説明的な特徴量について自動認識に置き換えた場合を意味する．*Has* については自動認識器が学習できなかったため，どの場合も *empty* としてモデルへと与えた．自動認識結果を用いる場合，どの場合についても *uttr+img+desc* と比較して大きく性能が低下している．*Uttr+img+desc* では正しい気の利いた行動を選択できているが，*uttr+img+desc^* では分類に重要なオブジェクトの認識を誤っておりかつ誤った行動を選択している例を図 4 に示す．どちらの例においても，気の利いた行動に関連したオブジェクトであるリモコンやゴミが認識できていない．これらのオブジェクトはデータセット中における出現回数が低いオブジェクトであった．図 4 の例のように *uttr+img+desc^* が動作の対象となるオブジェクトを認識できておらず，行動選択も誤ったケースは全誤り中 67.54% であった．

また *uttr^+img+desc^* と *uttr^* または *uttr^+img* との性能を比較した場合，R@5, MRR については有意に性能が向上しているものの，accuracy は変化していないことがわかった．しかし *uttr+img+desc^* と *uttr+img* との性能を比較した場合，どの指標についても有意に性能が向上しており，説明的な特徴量のみを自動認識した場合は行動選択の精度が向上することが判明した．

表 2 自動特徴量認識の結果

特徴量	WER	MER	WIL
uttr	27.46	25.75	32.8097
特徴量	Accuracy	-	-
position	98.25	-	-
pose	98.00	-	-
特徴量	Precision	Recall	F1 (%)
coffee table	91.36	78.55	84.46
dining table	86.02	87.54	86.76
kitchen	91.54	72.49	80.72

表 3 自動認識した特徴量による分類結果．対応のある T 検定で有意差を検定した．†† は $p < 0.01$ を，† は $p < 0.05$ を意味する．

モデル	Acc. (%)	R@5 (%)	MRR
uttr	27.02	53.85	0.4054
*uttr^*	22.10	43.80	0.3391
uttr+img	27.23	54.50	0.4064
uttr^+img	21.30	44.53	0.3369
uttr+img+desc	63.58	87.12	0.7417
uttr^+img+desc	††57.33	††83.93	††0.6921
*uttr+img+desc^*	††30.92	††61.80	††0.4577
*uttr^+img+desc^*	††22.20	††51.12	††0.3709

5 おわりに

本研究では，ユーザの要求発話が曖昧である場合に周囲の状況を理解しつつロボットの気の利いた行動を選択する分類モデルを構築し，様々な特徴量の与え方について比較評価を行った．具体的にはユーザ発話および周囲の状況を自動で認識した結果を与えた場合の分類性能を評価した．その結果，自動認識結果であっても周囲の状況に関する説明的特徴量を与えることは気の利いた行動の識別に効果的であり，事前学習モデルを単純に用いる場合よりも性能が向上することが示唆された．今後は周囲の状況を認識するモデルの精度を向上させ，実際のロボット上で動作する行動選択モデルとして実装する．

謝辞

本研究は理研の大学院生リサーチ・アソシエイト制度の下での成果である．本研究の一部は科研費 (22H03654) の支援を受けた．

参考文献

[1] Ryan Blake Jackson and Tom Williams. Enabling morally sensitive robotic clarification requests. *J. Hum.-Robot Interact.*, Vol. 11, No. 2, mar 2022.

[2] Felix Gervits, Antonio Roque, Gordon Briggs, Matthias Scheutz, and Matthew Marge. How should agents ask questions for situated learning? an annotated dialogue corpus. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 353–359, Singapore and Online, July 2021. Association for Computational Linguistics.

[3] Koichiro Yoshino, Yu Suzuki, and Satoshi Nakamura. Information navigation system with discovering user interests. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 356–359, Saarbrücken, Germany, August 2017. Association for Computational Linguistics.

[4] Shohei Tanaka, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura. ARTA: Collection and classification of ambiguous requests and thoughtful actions. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 77–88, Singapore and Online, July 2021. Association for Computational Linguistics.

[5] Robert S. Taylor. The process of asking questions. *American Documentation*, pp. 391–396, 1962.

[6] 田中翔平, 湯口彰重, 河野誠也, 中村哲, 吉野幸一郎. 気の利いた家庭内ロボット開発のための曖昧なユーザ要求と周囲の状況の収集. 情報処理学会 第 253 回自然言語処理研究会 (SIGNL), 2022.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. In *arXiv*, 2019.

[8] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 09–15 Jun 2019.

[9] Andrew Morris, Viktoria Maier, and Phil Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. 10 2004.

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

[11] Daisuke Kawahara. Kawahara Lab. RoBERTa (https://huggingface.co/nlp-waseda/roberta-base-japanese). 2021.

[12] Hakan Cevikalp, Burak Benligiray, and Omer Nezih Gerek. Semi-supervised robust deep neural networks for multi-label image classification. *Pattern Recognition*, Vol. 100, p. 107164, 2020.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

付録

実験設定

モデルの実装には PyTorch [10], 日本語 Wikipedia および CC-100 で事前学習された RoBERTa [11] を用いた．また収集したロボットの一人称視点を想定した動画からクリップされた画像を事前学習された EfficientNet-B0 で特徴量ベクトルへと変換した．RoBERTa および EfficientNet のパラメータもモデルの学習を通じてファインチューニングした．モデルの学習には hinge loss [12, 4] を用い，パラメータの最適化には Adam [13] を使用し，学習率は 1e−5 とした．

評価指標

R@5 は分類モデルが出力した正解カテゴリの順位が上位 5 位以内に含まれている割合である．MRR ($0 < MRR \leq 1$) は次式の通り算出される．

$$MRR = \frac{1}{|U_{test}|} \sum_{i}^{|U_{test}|} \frac{1}{r_{x_i}}. \quad (1)$$

ここで r_{x_i} はユーザ発話 x_i に対応する正解カテゴリについて分類モデルが出力した順位を意味し，U_{test} はテストデータに含まれるユーザ発話の集合である．全ての指標について，数値が大きいほど分類モデルの性能が高いことを意味する．各モデルの性能は五分割交差検証にて算出し，各分割データについて 10 回実験を試行した．