Vol.2022-NL-253 No.5
2022/9/29

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

田中 翔平[1,2,a)]　湯口 彰重[2,1,b)]　河野 誠也[2,c)]　中村 哲[1,d)]　吉野 幸一郎[2,1,e)]

概要：人と協働する対話ロボットは，ユーザの要求に応じて適切なタスク行動を行うことが一般的である．しかしユーザの要求はしばしば顕在化されず，対話ロボットはそうした状況でも，周囲の状況を適切に読み取りユーザが必要とする行動を取ることが期待される．こうした気の利いた行動をとることができるロボットを実現するため，リビングやキッチンにおいてユーザの家事を補助するタスクを対象に，ユーザの発話と周囲の状況に対応する気の利いたロボットの行動からなるデータを構築した．データ構築の方法として，本研究では大きく分けて三段階の手順を踏んだ．まず “ペットボトルを持ってくる” など，ロボットがとることのできる気の利いた行動をあらかじめ定義し，それらの行動をとっているロボットの動画を収録した．次に収集した行動の動画をクラウドワーカーに視聴してもらい，どのような状況でロボットがその行動をとってくれたら気が利いていると思うかをテキストで入力してもらった．最後に収集した状況のテキストに基づき，ロボットが気の利いた行動をとる直前のユーザの発話が行われる状況に紐付けられた動画を収集した．一般にロボットの学習で用いることができるデータは収集コストが大きいため，本研究ではごく少数のデータを収集し，収集した画像から得られる説明的な特徴量についてのアノテーションを行った．構築した少数データセットを用いて気の利いた行動を選択するロボットを実現するため，ユーザの発話内容や画像の畳み込みのみを特徴量として用いる分類器や，説明的な特徴量も用いるマルチモーダルな分類器など，複数のベースラインモデルを構築した．構築したベースラインモデルの性能を比較したところ，単純に画像の畳み込みや事前学習モデルによる特徴量抽出を用いるよりも，人手で付与した画像特徴に関する説明的なアノテーション結果がより分類精度の向上に寄与し，画像から抽出する情報の種類が重要であることが示された．

キーワード：対話システム，家庭内ロボット，曖昧な要求，気の利いた行動，マルチモーダル

1. はじめに

人と協働する対話ロボットの研究において，従来のロボットの多くは，ユーザがロボットに対する要求を明示的に示す，もしくは曖昧な要求であっても問い直しによって明確化できることを仮定している [5], [6]．しかし実際には，ユーザの要求はしばしば曖昧であり，明示的な要求を示すことができない場合も多い [18], [19], [20], [24]．“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 [19], [20]．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．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，状況に応じて能動的に (プロアクティブに) 振る舞いを決定することが求められる．

このような気の利いた行動をロボットで実現しようとする場合，いくつかの問題が存在する．まず，ロボットはユーザの発話だけでなく，ユーザを取り巻く状況からユーザが必要とする支援行動を導かなければならない．つまり，言語だけではなく音声，画像などから得られる情報を統合的に用いることが求められる．既存のマルチモーダル情報を用いた対話ロボットや対話エージェントの研究は，画像の内容についての質問応答 [11], [21] や要求の解析 [5]，ま

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology, Japan
[2] 理化学研究所ガーディアンロボットプロジェクト
Guardian Robot Project, RIKEN, Japan
a) tanaka.shohei.tj7@is.naist.jp
b) akishige.yuguchi@riken.jp
c) seiya.kawano@riken.jp
d) s-nakamura@is.naist.jp
e) koichiro.yoshino@riken.jp

表 1　ユーザ要求 (クエリ) の曖昧さの段階 [19], [20]

段階	定義
Q1	ユーザ自身も言語化できていない要求
Q2	ユーザ脳内では言語されているが発話されていない要求
Q3	明確に発話された要求
Q4	システムが取り扱いやすいフォーマットに変換された要求

たそれをトリガーとした雑談 [4], [12], [13], [26] などが主であり，認識した状況から気の利いた行動を求めることは行っていない．

また，ロボットでマルチモーダルデータを利用しようとした場合，学習データの量の問題に直面する [7]．近年一般に用いられる機械学習手法は大規模な学習データを必要とするが，ロボットはその身体性に合わせてデータを収集する必要があることから，データの大規模化は容易ではない．またユーザを含めた現実世界について，ロボットの学習に有効な再現度をもってシミュレーションすることも難しい [23]．限られたデータから効果的に機械学習を行おうとする場合，ロボットはその少数データから適切にタスクに有効な特徴量の抽象度を学習しなければならない [25]．

本研究では状況に応じて気の利いたロボットの行動を選択するため，ロボットの一人称視点からの画像と対応する気の利いたロボット行動のデータセットを構築した．構築の方法論として，まずロボットがとることのできる行動をあらかじめ定義し，それらの行動に先行する状況をクラウドソーシングで収集する方法 [18] を用いた．このクラウドソーシングで入力された状況に相当する状況をマルチモーダルデータとして収録した．また，限られたデータから気の利いたロボットの行動を選択できるようにするため，マルチモーダル情報から得られる状況を説明するような情報をアノテーションした．複数のベースラインモデルの性能を比較した実験結果から，ごく少数のデータセットのみが学習データとして利用可能な場合でも，画像から得られる説明的な特徴量を適切に活用することで曖昧な要求に対して気の利いた行動を選択する精度が大きく向上することがわかった．また，既存の事前学習モデルを用いる場合でもこうした説明的な特徴量を追加で用いることは有効であることが明らかになった．

2. タスク設計とデータ構築

Taylor [19], [20] は，情報探索時のユーザ要求を対象として，その明確度に応じて表 1 のような 4 段階に分類している．これまで研究・実用化されたきた何らかのユーザの要求を取り扱うシステム [15], [22] やロボット [1] は，ユーザ発話の中に明確に要求が含まれていることを前提とし，Q3 から Q4 の変換を行おうとするものである．これに対し，今後の対話ロボットやシステムでは Q1 や Q2 のようなユーザが自身の要求を明確に言語化できていない状況に対しても，ロボットがイニシアチブを取りつつその場で必要な行動を行うことができるようになることが求められる [24]．本研究ではこうした状況を想定し，クラウドソーシングを用いてプロアクティブな支援が必要な状況の収集を行った．

図 1　気の利いた対話の例: ロボットはバナナを持ってきている．

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．本研究では hello robot 社より販売されている Stretch RE1 [*1][*2] をこうした補助行動を行うロボットと想定する．Stretch RE1 はカメラとロボットアームを備えている家庭用モバイルマニピュレータであり，ロボットアームの耐荷重は 1.5 kg である．以下，本研究で取り扱うタスクの厳密な定義およびデータ構築方法について述べる．

2.1 タスク設定

本研究で取り扱う対話は，そのすべてで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の画像，それに対応したロボットの気の利いた行動の三つ組が定義されている．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双方を勘案し，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．本研究で定義したロボットの行動カテゴリのリストを表 2 に示す．定

[*1] https://hello-robot.com/
[*2] https://spectrum.ieee.org/hello-robots-stretch-mobile-manipulator

表 2　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

義した行動カテゴリは全部で 40 種類である．例えば図 1 のユーザ発話と状況が入力された場合，“バナナを持ってくる” という行動カテゴリを 40 種類のカテゴリの中から選択できれば，ロボットは気の利いた行動を選択できたとみなす．

2.2　ユーザの曖昧な要求とロボットの気の利いた行動の収集

図 1 のような対話を収集する方法として，2 人のワーカーにユーザ役とロボットを操作する役に分かれてインタラクションを行ってもらう WOZ 対話 [2], [8] が考えられる．しかし，意図の曖昧なユーザ要求に対して適切な気の利いた行動を選択することは，人間にとっても難しいタスクであり，一般的な WOZ 対話では期待するユーザの要求とロボットの行動のペアを収集することが難しい [18]．そこで本研究では，まずロボットがとることのできる気の利いた行動をあらかじめ定義した．そして定義したロボットの行動に対して，ユーザの先行発話および室内の状況を多数のクラウドワーカーに入力してもらうことで状況-行動ペアを収集した．例えば，グラスを持ってくるという行動に対応させて，“お酒を飲もうと話していて準備をしているときに，足らないグラスを探して “あれ？もう一個足りない” と言った．” などユーザ発話に加えてユーザが置かれた状況を詳細に説明してもらった．入力前にロボットが定義された行動をとっているビデオを教示することで，対象のロボットの行動の理解を円滑にした．表 2 にて定義したロボットの行動カテゴリに基づき，日本語コーパスをクラウドソーシング[*3]を用いて収集した．

[*3] https://crowdworks.jp/

2.3　視覚的特徴量のアノテーション

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 2.2 節において収集したコーパスには “あれ？もう一個足りない” という発話が含まれている．これ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” という画像情報を参照することで紐づけが可能となる．そこで本研究では，収集したユーザ発話に紐付けられた状況をリビングやキッチンにおいて演じた動画を収集した．

また今回は，ある環境でのロボットの一人称視点のデータを用いることを想定するため，大量のデータを用意することが難しい．こうした少量データの利用においては，学習データの抽象化が重要である [25]．そこで，画像から得られる様々な抽象化レベルの情報を利用するため，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．アノテーションの詳細を次に述べる．

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．これらの特徴は人手で付与したが，今後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．すべての画像内には必ずコーヒーテーブル，ダイニングテーブル，キッチンが写り込んでいる．図 2 のアノテーション例より，発話に紐付けられた画像，画像から得られる特徴量が対象とする状況と気の利いたロボットの行動に合わせてアノテーションされていることがわかる．収集したコーパスの統計情報を表 3 に示す．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，本研究では 400 本のデータのみを収集した．これは一般的な対話コーパス [2], [18] と比較すると非常に少ない数である．

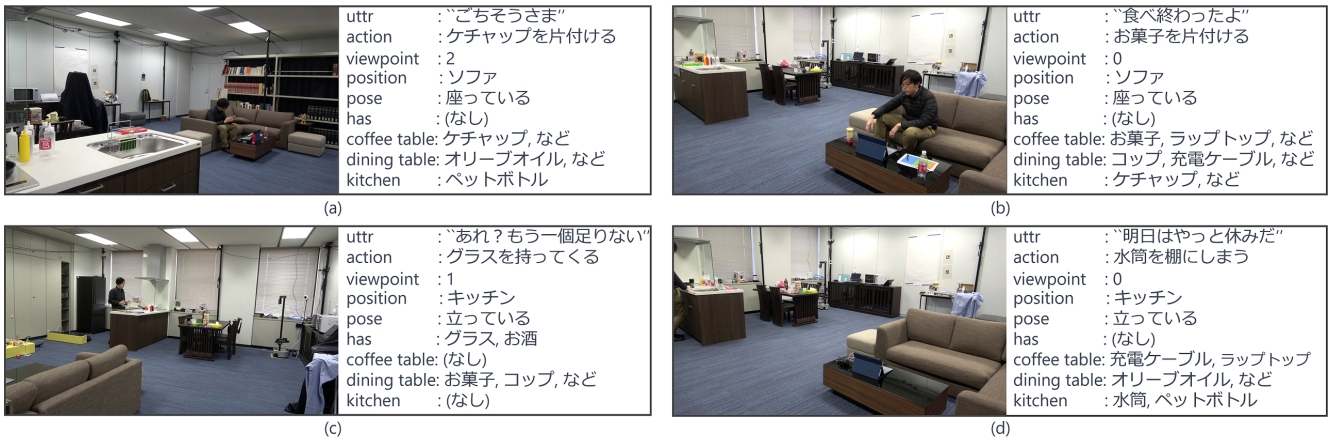

図 2 収集した対話の例

表 3 収集したコーパスの統計情報

対話数	400
平均発話長	11.59 (±4.94)
ユーザが持っているオブジェクト数	0.14 (±0.37)
コーヒーテーブル上のオブジェクト数	1.40 (±0.91)
ダイニングテーブル上のオブジェクト数	4.32 (±1.95)
キッチン台上のオブジェクト数	1.47 (±1.66)

3. 視覚的特徴量を用いた気の利いた行動のベースライン分類器

2 章で収集したデータを用い，実際に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．まずユーザ発話を入力とすることが考えられるが，2.2 節で示した通り，多くのケースでは曖昧なユーザ発話のみから適切な行動クラスを推定することは容易ではない．マルチモーダルな特徴量を利用する場合，既存の研究は事前学習モデルに基づく特徴量抽出を行う [17]．しかし，ロボットの行動クラス推定のように少ない学習データを利用しようとする場合，マルチモーダル情報から利用される情報はより抽象化される必要がある [25]．そこで今回のベースライン分類器では，各状況に付与された物体やユーザの状態などのアノテーションを入力として用い，これらが分類精度向上にどの程度寄与するか明らかにする．

図 3 に今回構築したベースライン分類器の概要を示す．マルチモーダル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 クラスから推定する．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 [14] 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ことで，この発話が出現する文脈が特徴量として利用できることを期待する．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 RoBERTa へ入力し [CLS] ベクトルを特徴量ベクトルとする．これらの特徴量は単語単位であるため，特徴ベクトルへの変換に用いるネットワークを単純な word embedding 層で代替する方法も考えられる．だが実際に word embedding 層で代替したモデルと RoBERTa を利用したモデルの性能を比較したところ，word embedding 層を用いるモデルの性能が有意に低下したため，本研究では単語単位の特徴量も RoBERTa を用いて特徴量ベクトルへと変換した．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 [17] に入力してベクトル化する．この画像特徴と発話特徴のみを用いた場合のベースライン分類器を uttr+image と定義する．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

3 章にて構築した，ユーザ発話を対応する応答のカテゴリへと分類するベースラインモデルを評価する．

4.1 実験設定

モデルの実装には PyTorch [16], 日本語 Wikipedia および CC-100 で事前学習された RoBERTa [9] を用いた．また収集したロボットの一人称視点の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 のパラメータはモデルの学習を通じてアップデートした．モデルの学習には hinge loss [3], [18] を用い，パラメータの最適化には Adam [10] を使用し，学習率は 1e−5 とした．

評価指標として，Accuracy (Acc.), Recall@5 (R@5), Mean Reciprocal Rank (MRR) を用いた．R@5 は，分

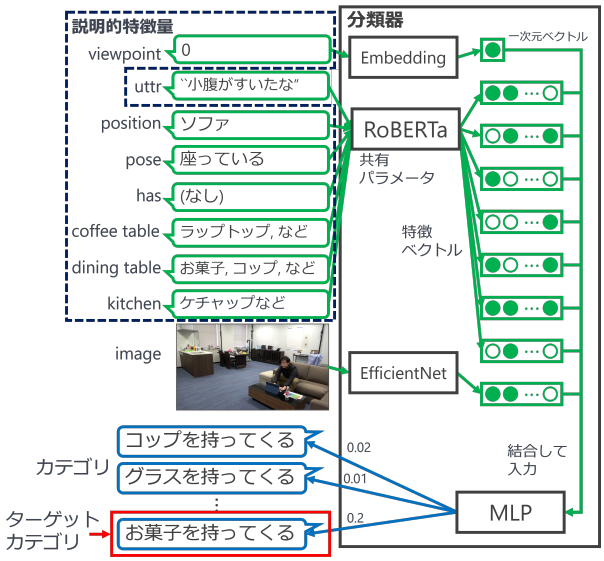

図 3　ベースライン分類器における特徴量の入力

表 4　50 回試行した分類結果の平均値．対応のある T 検定で有意差を検定した．†† は $p < 0.01$ を意味する．

モデル	Acc. (%)	R@5 (%)	MRR
uttr	27.02	53.85	0.4054
uttr+img	27.23	54.50	0.4064
uttr+img+desc	††63.58	††87.12	††0.7417

表 5　削除実験の分類結果．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+img	27.23	54.50	0.4064
uttr+img+desc	63.58	87.12	0.7417
w/o viewpoint	††60.08	†85.62	††0.7132
w/o user	64.12	86.75	0.7424
w/o object	††31.38	††63.28	††0.4680
w/o image	††61.17	†85.75	††0.7231

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．全ての指標について，値が高いほどモデルの性能が高いことを意味する．実験に用いる各モデルのパラメータは，検証データにおける損失関数の値が最小のものを使用した．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．

4.2　気の利いた行動の分類性能

各分類器の評価結果を評価した結果を表 4 に示す．*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) を入力とした場合を意味する．*Uttr+img+desc* と *uttr+img* の性能の有意差を対応のある T 検定で検定した．† は $p < 0.05$ で，†† は $p < 0.01$ で有意に性能が向上したことを意味する．どの指標についても，画像から得られる説明的な特徴量も用いた方が劇的に性能が向上していることがわかる．これは本研究で想定する実空間でのロボットデータを用いた分類を考える上で重要な示唆であり，こうした有効な特徴量ラベルの設定が必要である．

4.3　詳細分析

図 3 における説明的な特徴量をそれぞれ取り除いた場合の削除モデル性能を表 5 に示す．*User* はユーザに関する特徴量であり，図 3 における position, pose がそれに含まれる．*Object* は画像に写り込んでいる物体に関する特徴量であり，図 3 における has, coffee table, dining table, kitchen がそれに含まれる．各削除モデルについて，*uttr+img+desc* との性能の有意差を対応のある T 検定で検定した．表 5 より，*object* を取り除いた場合に最も性能が低下しており，*object* が状況に対して気の利いた行動を選択するために画像から得られる特徴量として最も重要であることがわかる．また *object* 以外の特徴量に関しては取り除いても性能があまり低下しないため，一見すると重要でない特徴量である．しかし *w/o object* と *uttr+img* の性能についても有意差を検証したところ，$p < 0.01$ で有意差があったため，*object* に関する特徴量が利用できない場合はこれらの特徴量も有用だと言える．*Object* に関する特徴量のアノテーションには *viewpoint* や *user* に関する特徴量と比較してコストが掛かるため，より低コストで分類精度を向上させたい場合，他の特徴量のみを利用することは有効な選択肢となる．

Uttr+img では気の利いた行動の選択を誤っているが，*uttr+img+desc* では適切に選択できている例を図 4 に示す．*Gold* は正解として紐付けられた気の利いた行動を意味する．各ケースについて最も重要な説明的特徴量である物体の各ラベルに与えられた RoBERTa の注視重み (Attention) を可視化した．多くのケースについて気の利いた行動に関連した物体に分類器の注視が当たっており，*uttr+img* では全く関係のない行動を選択していた状況において *uttr+img+desc* は適切な行動を選択できている．すなわち，説明的な特徴量も用いることで少ない学習データから適切に気の利いた行動の推定が可能であることがわかる．しかし，図 4 の (d) に示すとおり，モデルが適切な物体に注視を当てられていない場合も存在する．よって，ユーザを取り巻く状況と気の利いた行動の因果関係を因果推論 [27] の枠組みで明示的に取り扱うなど，より正確に状況を理解できるようなモデルアーキテクチャを開発する必要がある．

5.　おわりに

本研究は，ユーザの要求発話が曖昧である場合に，周囲の状況を見つつ気の利いた行動を行うロボットの行動選択

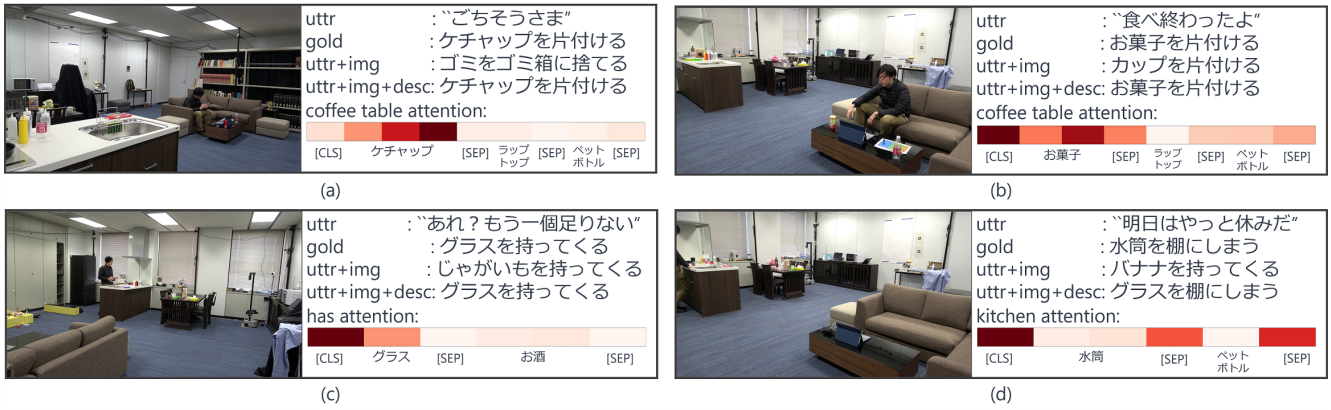

図 4 分類器が視覚的特徴量を活用した事例: 暗色は強いアテンションを意味する．*Uttr+img+desc* はケチャップ，お菓子，グラスなど (a), (b), (c) それぞれの事例において重要なオブジェクトにアテンションを当てている．しかし (d) では水筒にアテンションが当たっていない．

を学習させるためのデータセットを構築したものである．具体的には，あるロボットのユーザ補助行動に先行する生活空間での状況をクラウドソーシングによって収集し，入力された状況に対応する生活空間でのシナリオを実際に収録した．この際，ロボットは一人称視点でユーザの状況を観察し，現在の状況から行うべき気の利いた行動の推定を行う．この推定を適切に行うには，画像から分類に寄与する説明的特徴量を抽出することが重要であり，実際にアノテーションされた説明的な特徴量が分類に寄与することを示した．今後は実際に本システムをロボットに搭載し，説明的特徴量についても推定を自動化することで，人間の生活空間で協調的に気を利かせて動作することができるロボットシステムの構築を目指す．

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] Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Ho, D., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jang, E., Ruano, R. J., Jeffrey, K., Jesmonth, S., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Lee, K.-H., Levine, S., Lu, Y., Luu, L., Parada, C., Pastor, P., Quiambao, J., Rao, K., Rettinghouse, J., Reyes, D., Sermanet, P., Sievers, N., Tan, C., Toshev, A., Vanhoucke, V., Xia, F., Xiao, T., Xu, P., Xu, S., Yan, M. and Zeng, A.: Do As I Can and Not As I Say: Grounding Language in Robotic Affordances, *arXiv preprint arXiv:2204.01691* (2022).

[2] Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O. and Gašić, M.: MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Association for Computational Linguistics, pp. 5016–5026 (2018).

[3] Cevikalp, H., Benligiray, B. and Gerek, O. N.: Semi-supervised robust deep neural networks for multi-label image classification, *Pattern Recognition*, Vol. 100, p. 107164 (2020).

[4] Chiba, Y. and Higashinaka, R.: Dialogue Situation Recognition for Everyday Conversation Using Multimodal Information, *Proceedings of INTERSPEECH 2021 (INTERSPEECH)*, pp. 241–245 (online), DOI: 10.21437/Interspeech.2021-171 (2021).

[5] Gervits, F., Roque, A., Briggs, G., Scheutz, M. and Marge, M.: How Should Agents Ask Questions For Situated Learning? An Annotated Dialogue Corpus, *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Singapore and Online, Association for Computational Linguistics, pp. 353–359 (online), available from ⟨https://aclanthology.org/2021.sigdial-1.37⟩ (2021).

[6] Jackson, R. B. and Williams, T.: Enabling Morally Sensitive Robotic Clarification Requests, *J. Hum.-Robot Interact.*, Vol. 11, No. 2 (online), DOI: 10.1145/3503795 (2022).

[7] James, S., Bloesch, M. and Davison, A. J.: Task-Embedded Control Networks for Few-Shot Imitation Learning, *CoRR*, Vol. abs/1810.03237 (online), available from ⟨http://arxiv.org/abs/1810.03237⟩ (2018).

[8] Kang, D., Balakrishnan, A., Shah, P., Crook, P., Boureau, Y.-L. and Weston, J.: Recommendation as a Communication Game: Self-Supervised Bot-Play for Goal-oriented Dialogue, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Association for Computational Linguistics, pp. 1951–1961 (2019).

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

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

[11] Le, H., Sankar, C., Moon, S., Beirami, A., Gerami-

fard, A. and Kottur, S.: DVD: A Diagnostic Dataset for Multi-step Reasoning in Video Grounded Dialogue, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, Association for Computational Linguistics, pp. 5651–5665 (online), DOI: 10.18653/v1/2021.acl-long.439 (2021).

[12] Lee, N., Shin, S., Choo, J., Choi, H.-J. and Myaeng, S.-H.: Constructing Multi-Modal Dialogue Dataset by Replacing Text with Semantically Relevant Images, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Online, Association for Computational Linguistics, pp. 897–906 (online), DOI: 10.18653/v1/2021.acl-short.113 (2021).

[13] Liang, Z., Hu, H., Xu, C., Tao, C., Geng, X., Chen, Y., Liang, F. and Jiang, D.: Maria: A Visual Experience Powered Conversational Agent, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, Association for Computational Linguistics, pp. 5596–5611 (online), DOI: 10.18653/v1/2021.acl-long.435 (2021).

[14] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019).

[15] Madotto, A., Wu, C.-S. and Fung, P.: Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1468–1478 (2018).

[16] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, *Advances in Neural Information Processing Systems 32* (Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R., eds.), Curran Associates, Inc., pp. 8024–8035 (2019).

[17] Tan, M. and Le, Q. V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks (2020).

[18] Tanaka, S., Yoshino, K., Sudoh, K. and Nakamura, S.: ARTA: Collection and Classification of Ambiguous Requests and Thoughtful Actions, *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Singapore and Online, Association for Computational Linguistics, pp. 77–88 (online), available from ⟨https://aclanthology.org/2021.sigdial-1.9⟩ (2021).

[19] Taylor, R. S.: The process of asking questions, *American Documentation*, pp. 391–396 (1962).

[20] Taylor, R. S.: Question-Negotiation and Information Seeking in Libraries, *College & Research Libraries*, Vol. 29, No. 3, pp. 178–194 (1968).

[21] Testoni, A. and Bernardi, R.: Looking for Confirmations: An Effective and Human-Like Visual Dialogue Strategy, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, Association for Computational Linguistics, pp. 9330–9338 (online), DOI: 10.18653/v1/2021.emnlp-main.736 (2021).

[22] Vanzo, A., Bastianelli, E. and Lemon, O.: Hierarchical Multi-Task Natural Language Understanding for Cross-domain Conversational AI: HERMIT NLU, *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Stockholm, Sweden, Association for Computational Linguistics, pp. 254–263 (2019).

[23] Weibel, J., Patten, T. and Vincze, M.: Addressing the Sim2Real Gap in Robotic 3D Object Classification, *CoRR*, Vol. abs/1910.12585 (online), available from ⟨http://arxiv.org/abs/1910.12585⟩ (2019).

[24] Yoshino, K., Suzuki, Y. and Nakamura, S.: Information Navigation System with Discovering User Interests, *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Saarbrücken, Germany, Association for Computational Linguistics, pp. 356–359 (2017).

[25] Yoshino, K., Wakimoto, K., Nishimura, Y. and Nakamura, S.: Caption Generation of Robot Behaviors based on Unsupervised Learning of Action Segments, *CoRR*, Vol. abs/2003.10066 (online), available from ⟨https://arxiv.org/abs/2003.10066⟩ (2020).

[26] Zang, X., Liu, L., Wang, M., Song, Y., Zhang, H. and Chen, J.: PhotoChat: A Human-Human Dialogue Dataset With Photo Sharing Behavior For Joint Image-Text Modeling, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, Association for Computational Linguistics, pp. 6142–6152 (online), DOI: 10.18653/v1/2021.acl-long.479 (2021).

[27] 清丸寛一，植田暢大，児玉貴志，田中　佑，岸本裕大，田中リベカ，河原大輔，黒橋禎夫：因果関係グラフ: 構造的言語処理に基づくイベントの原因・結果・解決策の集約，言語処理学会第 26 回年次大会 発表論文集 (ANLP)，pp. 1125–1128 (2020).