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[12], 湯口彰重[21], 河野誠也[2], 中村哲[1], 吉野幸一郎[21]

[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

既存の家庭内口ボット

ユーザ

ユーザ発話は
特定の機能に
対する要求では
ない

あれ？もう
一個足りない

…

家庭内口ボット

要求の内容が
理解できない…

ユーザ要求に対して自身の機能にもとづいて行動する

ユーザ要求が明確であることを仮定している

ユーザ要求が曖昧な場合適切な行動を選択することができない

Copyright © 2022 Shohei Tanaka All Rights Reserved

研究目標：気の利いた家庭内ロボット

家庭内ロボットがユーザの潜在的な要求に
紐付けられる気の利いた行動をとる

気心の知れた人間同士は相手が必要としそうな
「気の利いた行動」をとることができる

曖昧な要求に対して気の利いた行動をとることが
できる気の利いた家庭内ロボットの開発を目指す

Copyright © 2022 Shohei Tanaka All Rights Reserved

気の利いた (reflective) とは？

Reactive

behave in response to what happens to them, rather than deciding in advance how they want to behave [Collins]

ユーザに言われたことをする行動 (従来の家庭内ロボット)

Proactive

intended to cause changes, rather than just reacting to change [Collins]

ロボットが主導権を持ってユーザに提案する行動 (広告や営業)

Reflective

thinking deeply about something [Collins]

ユーザのことを慮ったユーザに言われた以上の内容を含む行動

Copyright © 2022 Shohei Tanaka All Rights Reserved

タスク設定

家庭内でロボットがユーザを支援する状況を想定

ユーザ

あれ？もう
一個足りない

ユーザ発話 (要求)

ロボットの一人称
視点の画像

家庭内ロボット

ユーザはグラス
を持っている

ユーザ発話と画像に
基づいて判断

グラスは
いりますか？

気の利いた行動

入力：ユーザの曖昧な要求、ユーザや周囲の状況を
表すロボットの一人称視点の画像

出力：ロボットの気の利いた行動

Copyright © 2022 Shohei Tanaka All Rights Reserved

関連研究: SayCan

Google が発表した家庭内ロボット [https://say-can.github.io/]

「テーブルを拭いて」などの要求に対して言語モデルと行動評価関数を用いてロボットの行動系列を導く

スポンジを見つける -> スポンジを持ってくる -> テーブルを拭く

ユーザがロボットに対して明確に行動系列を命令しないというタスク設定は本研究と類似している

本研究はより曖昧な要求 (そもそも命令しない) に焦点を絞っている

Copyright © 2022 Shohei Tanaka All Rights Reserved

コーパスの収集方法

従来の WOZ (Wizard of OZ) 法の問題：

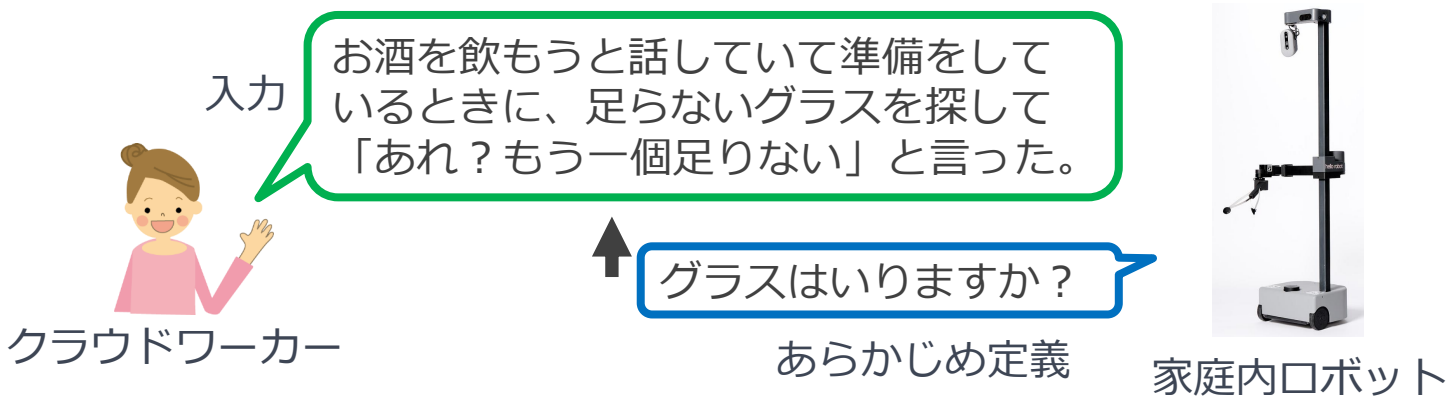
気の利いた行動をとることは人間にとっても難しい

ロボットの行動は自身の機能によって制限されている

↓

ロボットの行動はあらかじめ定められていた方が
コーパスの質を担保しやすい

**ワーカーにロボットが定義された行動をとっている
ビデオを教示し、その行動が気が利いていると
みなせるような先行発話と状況を入力してもらう**



Copyright © 2022 Shohei Tanaka All Rights Reserved

使用するロボットと行動カテゴリ

ロボットは Stretch RE1 を使用

- 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
- ロボットアームの耐荷重は 1.5 kg

Stretch RE1

ロボットの行動カテゴリは Stretch RE1 が実行可能な行動に基づいて定義

カテゴリ	カテゴリ数
グラスを持ってくる、コップを片付ける、ゴミをゴミ箱に捨てる、など	40

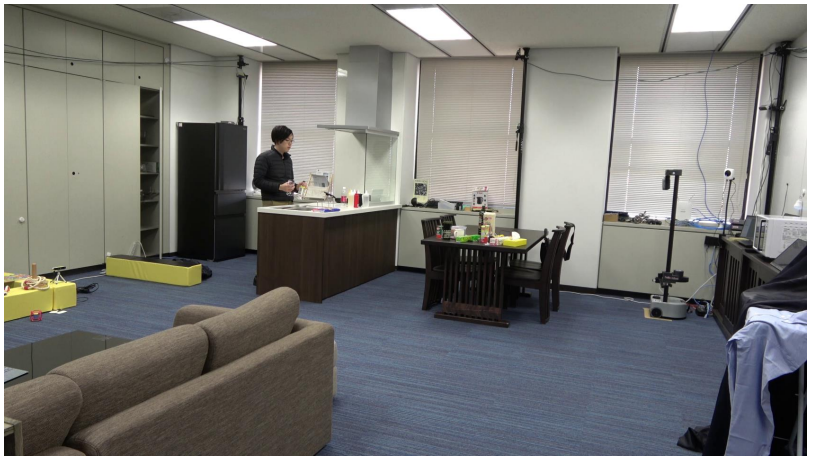
カテゴリごとに10対話、合計400対話を収集

Copyright © 2022 Shohei Tanaka All Rights Reserved

周囲の状況を表す動画の収集

ユーザの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

あれ？もう
一個足りない



ユーザ発話の内容のみから
「グラスを持ってくる」
行動を選択することは困難

収集したシナリオに従ってユーザの周囲の状況を表すロボットの一人称視点の動画を収集

動画の最後のフレームを状況を表す代表画像としてクリップ

Copyright © 2022 Shohei Tanaka All Rights Reserved

収集した動画

Copyright © 2022 Shohei Tanaka All Rights Reserved

説明的ラベルのアノテーション

400本という少数データで単純に事前学習モデルをファインチューニングすることは難しい

ユーザの周囲の状況を正確に認識するには状況を正確に記述したラベルでモデルを学習させる必要がある

↓

少数データを効率よく学習に利用するため、動画からクリップされた画像に**状況を説明するラベルを付与**

- ロボットの位置 (視点)
- ユーザの位置、姿勢、持っているオブジェクト
- コーヒーテーブル、ダイニングテーブル、キッチンにそれぞれあるオブジェクト

Copyright © 2022 Shohei Tanaka All Rights Reserved

説明的ラベルを付与されたデータの例

ロボットの視点番号、ユーザの位置、ユーザが持っているもの、
キッチンにあるもの、など状況を説明するラベルを人手で付与

Copyright © 2022 Shohei Tanaka All Rights Reserved

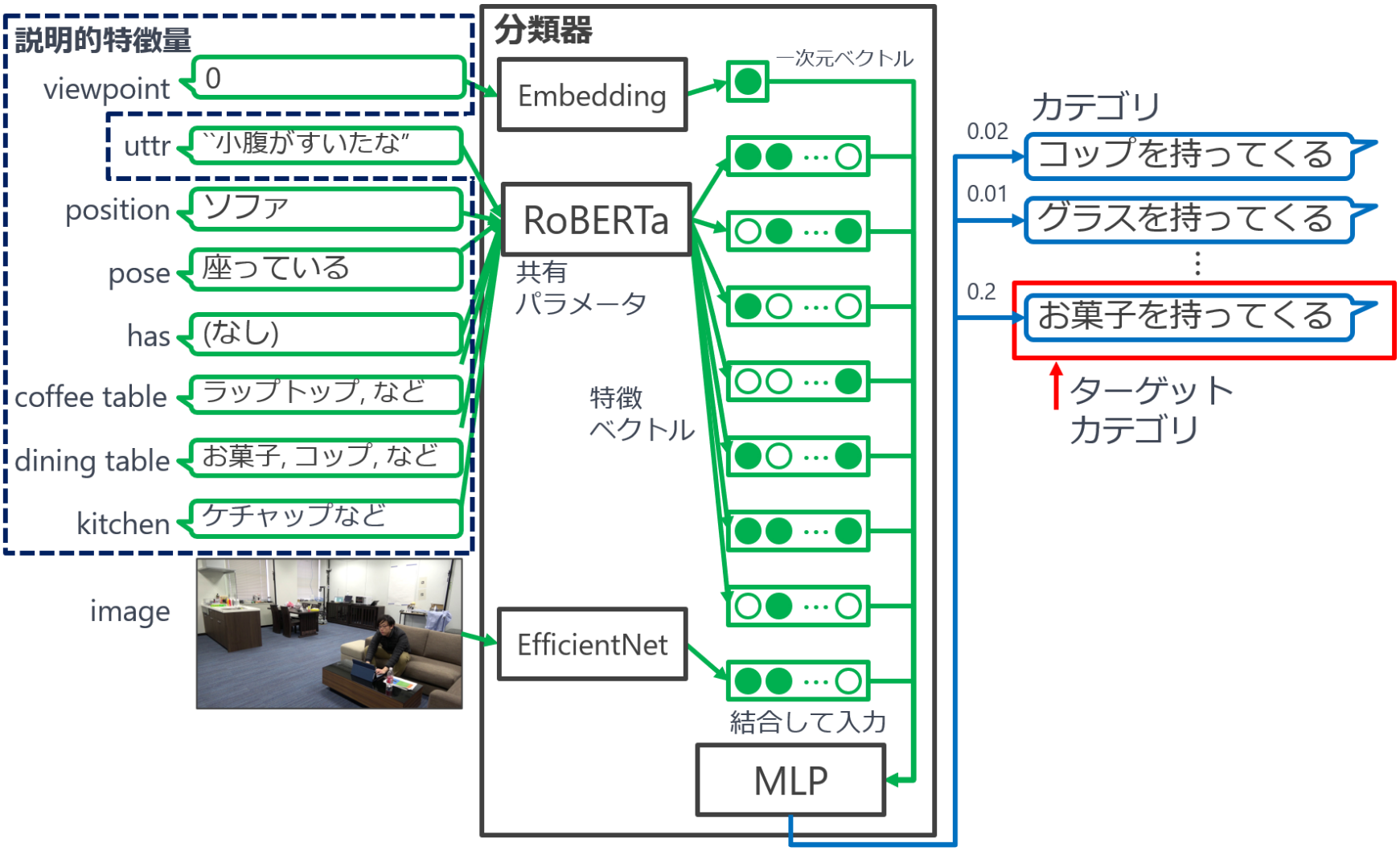
収集したデータの統計情報

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)

収集した対話数は400本であり、一般的なテキスト対話コーパスと比較してごく少数
ユーザが持っているオブジェクトは他のオブジェクトと比較して非常に少ない

Copyright © 2022 Shohei Tanaka All Rights Reserved

気の利いた行動のベースライン分類器



Copyright © 2022 Shohei Tanaka All Rights Reserved

実験設定

事前学習された日本語 RoBERTa [Kawahara, 2021], EfficientNet [Tan, 2020] を用いる

400本のデータを五分割交差検証にて評価

各分割データについて10回実験を試行

評価指標

- Accuracy
- R@5: 正解カテゴリが上位5位以内に含まれている割合
- Mean Reciprocal Rank (0 < MRR <= 1):

$$MRR = \frac{1}{|U|} \sum_{u \in U} \frac{1}{rank_u}$$

$rank_u$: 正解カテゴリの順位

$|U|$: テストデータ中のユーザ要求

Copyright © 2022 Shohei Tanaka All Rights Reserved

気の利いた行動の分類性能

** は $p < 0.01$ を意味する

モデル	Accuracy (%)	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**

uttr: ユーザ発話のみを入力とした場合
uttr+img: ユーザ発話と画像を入力とした場合
uttr+img+desc: ユーザ発話と画像に加え説明的な特徴量も入力とした場合

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利用したほうが性能が劇的に向上

こうした気の利いた行動の選択に有効な特徴量ラベルを画像から抽出することが重要である

Copyright © 2022 Shohei Tanaka All Rights Reserved

削除実験

** は p < 0.01 を、* は p < 0.05 を意味する

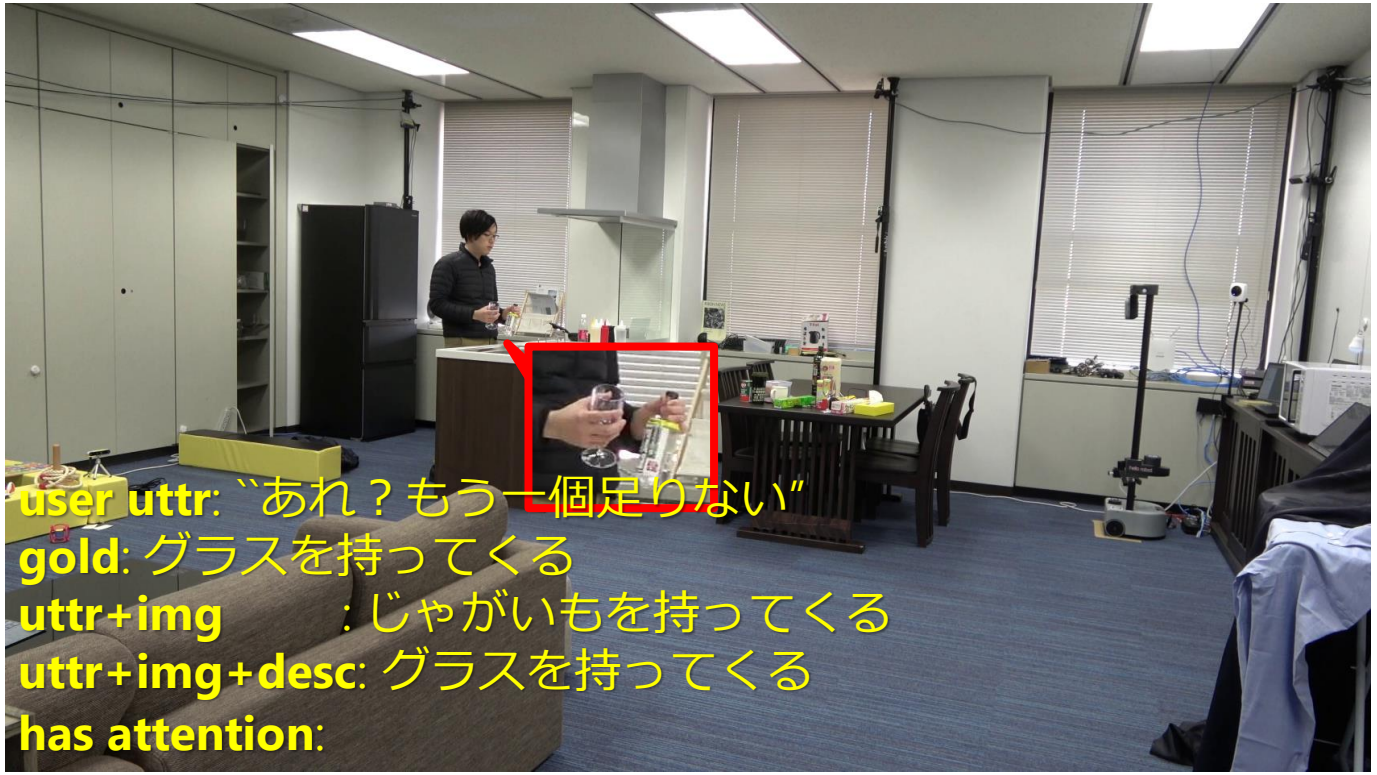
モデル	Accuracy (%)	R@5 (%)	MRR
uttr+img	27.23	54.50	0.4064
uttr+img+desc	**63.58**	**87.12**	**0.7417**
w/o viewpoint	**60.08	*85.62	**0.7132
w/o user	64.12	86.75	0.7424
w/o object	**31.38	**63.28	**0.4680
w/o image	**61.17	*85.75	**0.7231

user: ユーザに関する特徴量 (position, pose)
object: 画像に写っているオブジェクトに関する特徴量 (has, coffee table, dining table, kitchen)

- w/o object で最も性能が低下
- uttr+img と w/o object にも有意差あり
- オブジェクトのアノテーションにはコストが掛かるため、他の特徴量のみを利用することも有用

Copyright © 2022 Shohei Tanaka All Rights Reserved

注視重みの可視化 (1)



説明的な特徴量を利用することで、ユーザ発話と画像のみでは選択できなかった気の利いた行動を選択可能

Copyright © 2022 Shohei Tanaka All Rights Reserved

注視重みの可視化 (2)

説明的な特徴量を利用することで、ユーザ発話と画像のみでは選択できなかった気の利いた行動を選択可能

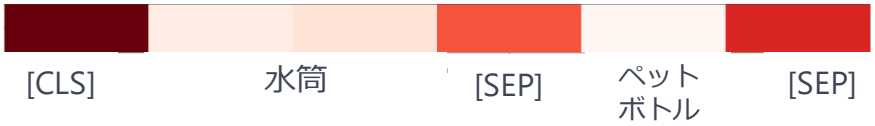
Copyright © 2022 Shohei Tanaka All Rights Reserved

注視重みの可視化 (3)

説明的な特徴量を利用することで、ユーザ発話と画像のみでは選択できなかった気の利いた行動を選択可能

Copyright © 2022 Shohei Tanaka All Rights Reserved

注視重みの可視化 (4)



適切に注視を当てられていない場合もある
因果推論などの枠組みでより正確に状況と気の利いた
行動を結びつけるアーキテクチャが必要

Copyright © 2022 Shohei Tanaka All Rights Reserved

おわりに

まとめ

ユーザの曖昧な要求と周囲の状況を表す画像、家庭内ロボットの気の利いた行動を含んだコーパスを収集

ロボットの一人称視点を想定した画像から得られる説明的な特徴量をアノテーション

ユーザの要求と周囲の状況にもとづいてロボットの行動を選択するベースラインモデルを構築

説明的な特徴量を用いることが行動選択の精度を向上させることを証明

今後の課題

実際に本システムをロボットに搭載するため、説明的な特徴量についての推定を自動化

Copyright © 2022 Shohei Tanaka All Rights Reserved