H12-1

# 実世界における総合的参照解析を目的としたマルチモーダル対話データセットの構築

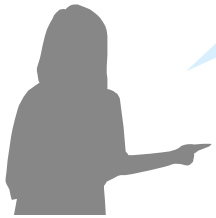
植田 暢大[1,2]　波部 英子[2]　湯口 彰重[2,3]　河野 誠也[2]
川西 康友[2,3]　黒橋 禎夫[1,2]　吉野 幸一郎[2,3]

[1]京都大学　[2]理化学研究所　[3]奈良先端科学技術大学院大学

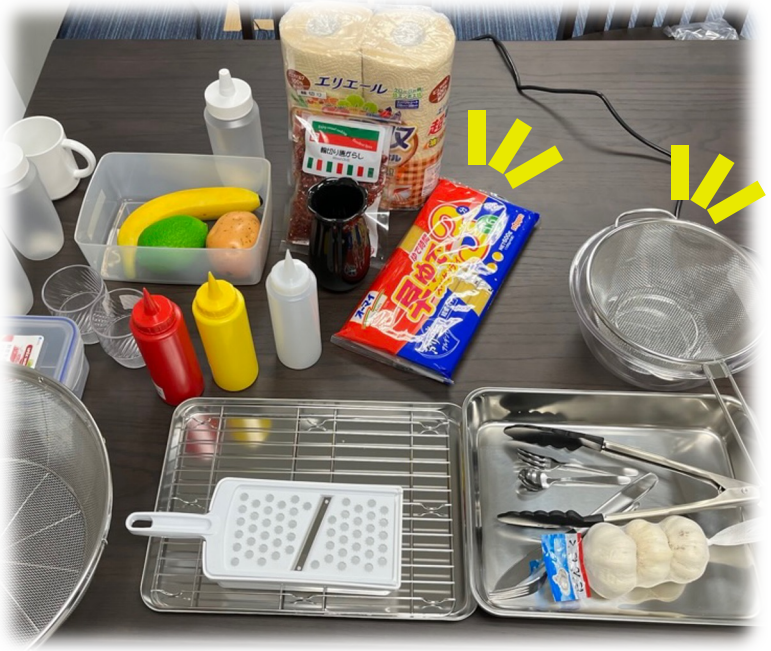
言語処理学会第29回年次大会 2023/3/16

# 研究背景：実世界対話における物体への参照

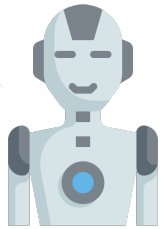
- 視覚情報が共有される実世界での対話では，**物体への参照**が頻出
- 参照先物体の特定はコンピュータによる発話理解に不可欠



そこの**ザル**を持ってきて




小さい方のザルですね
何を作るんですか？



ペペロンチーノね
あと**封**を切って準備して

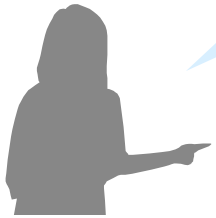
分かりました
パスタの封を切っておきます

お願いね

実世界対話において，発話中フレーズの参照先を物体を含め特定するシステムの構築を目指す

# 提案タスク：マルチモーダル参照解析

タスク設定：発話の書き起こしテキストと視覚シーンが与えられたとき，テキスト中の表現の参照先をフレーズおよび物体矩形として特定



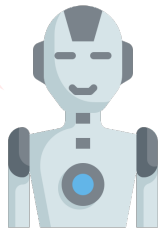
そこの**ザルを持ってきて**

ペペロンチーノね
あと**封を切って**準備して

お願いね


(ロボットが) (主人に) ザルを持ってくる

小さい方のザルですね
何を作るんですか？



ゼロ照応

(ロボットが) (パスタの) 封を切る

分かりました
パスタの封を切っておきます

# 関連研究: SIMMC 2.0 [Kottur+, EMNLP20]

Situated and Interactive Multimodal Conversations dataset

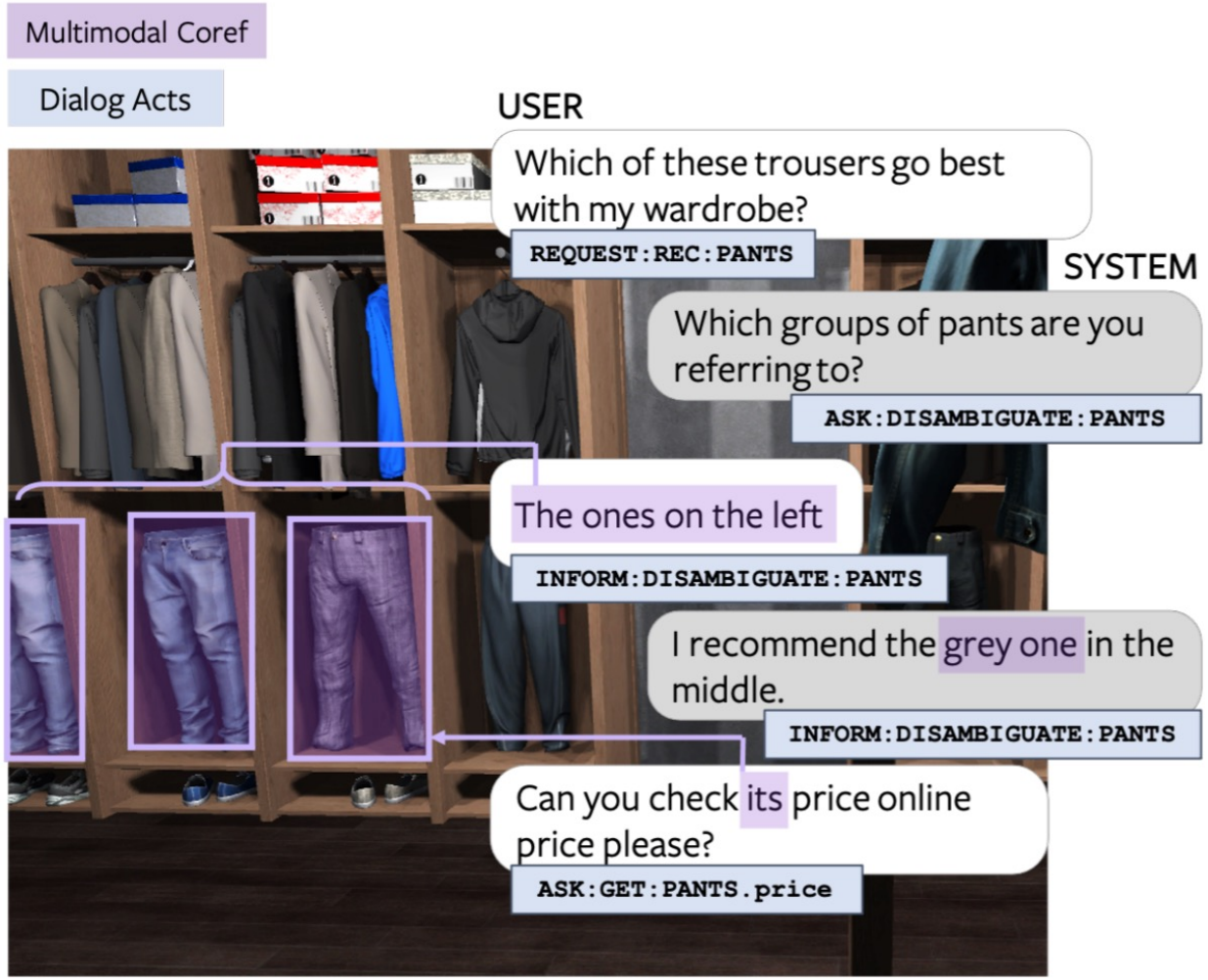
- 含まれるデータ: CG画像と対話テキスト

  本研究：1人称視点実世界動画

- アノテーション：対話行為，物体矩形，名詞句と物体矩形の対応

  本研究：ゼロ照応も含む

- 規模：1.5k枚の画像・11k対話




(a) SIMMC 2.0: Cluttered, closer-to-real-world multimodal contexts

# 研究の流れ

1. 対話シナリオ収集
- クラウドソーシング
- 30 対話 x 3 場面
- 場面: リビング, ダイニング, キッチン

2. 対話収録
- 主人とロボットを想定した人対人対話
- 音声および1・3人称視点動画

3. テキストおよび画像アノテーション
- 対話音声を書き起こし
- テキスト間照応をタグ付け
- テキスト・物体間参照をタグ付け

4. マルチモーダル参照解析システムの構築
- 既存の視覚・言語データセットや照応データセットを活用

# 1. 対話シナリオ収集：例

主人とそのお手伝いロボットの対話を想定したシナリオをクラウドソーシングで収集

| 話者 | 発話 |
|---|---|
| 主人 | 人形を梱包したいんだけど、紙をシュレッダーにかけて緩衝材を作ってくれない？ |
| ロボット | 分かりました。どの紙を使ったらいいですか？ |
| 主人 | （部屋の隅の雑誌の山を指差して）あそこから適当に使って。その段ボール箱に一杯分くらい。 |
| ロボット | （古雑誌を黙々とシュレッダーにかける）このくらいでよろしいですか？ |
| 主人 | （段ボールの中を確認しながら）うん、大丈夫。そしたら、そこにあるプチプチを持ってきてもらえる？ |
| ロボット | 分かりました。（プチプチを渡す）ついでに、それも包みましょうか？ |
| 主人 | ううん、壊れ物だから自分で包むよ。（人形を包む）テープを持ってきて。 |
| ロボット | テープはどこにありますか？ |
| 主人 | （棚を指差して）たぶん右の戸棚に入っていると思う。（人形を箱に入れる） |
| ロボット | （棚に向かい、クラフトテープを掴み）これでよろしいですか？ |
| 主人 | うん。人形は詰めたから、あとはテープで蓋を閉じて、玄関先まで運んでおいて。 |
| ロボット | 分かりました。（段ボールをテープで閉じる）作業が終わりましたので、箱を玄関先に置いてきます。 |

**10—16 発話**

**括弧内テキスト：ワーカーによる ト書き**

# 2. 対話収録：実験設定

- 対話設定
  - シナリオに基づいた主人役とロボット役の人対人対話
  - リビング，ダイニング，キッチンを模した環境で収録
- 設備
  - GoPro（1人称視点動画撮影用）
  - ピンマイク（対話音声録音用）
  - 定点カメラ x 4（3人称視点動画撮影用 ）
    - 1人称視点で起こりうる参照物体のオクルージョンをカバーする目的で補助的に使用
- シナリオ数

| | リビング | ダイニング | キッチン |
|---|---|---|---|
| 現在 | 19 | 15 | 17 |
| 目標 | 30 | 30 | 30 |

# 2. 対話収録：例

1人称視点動画（ロボット役視点）

3人称視点動画 x 4

# 3. テキストおよび画像アノテーション：概要

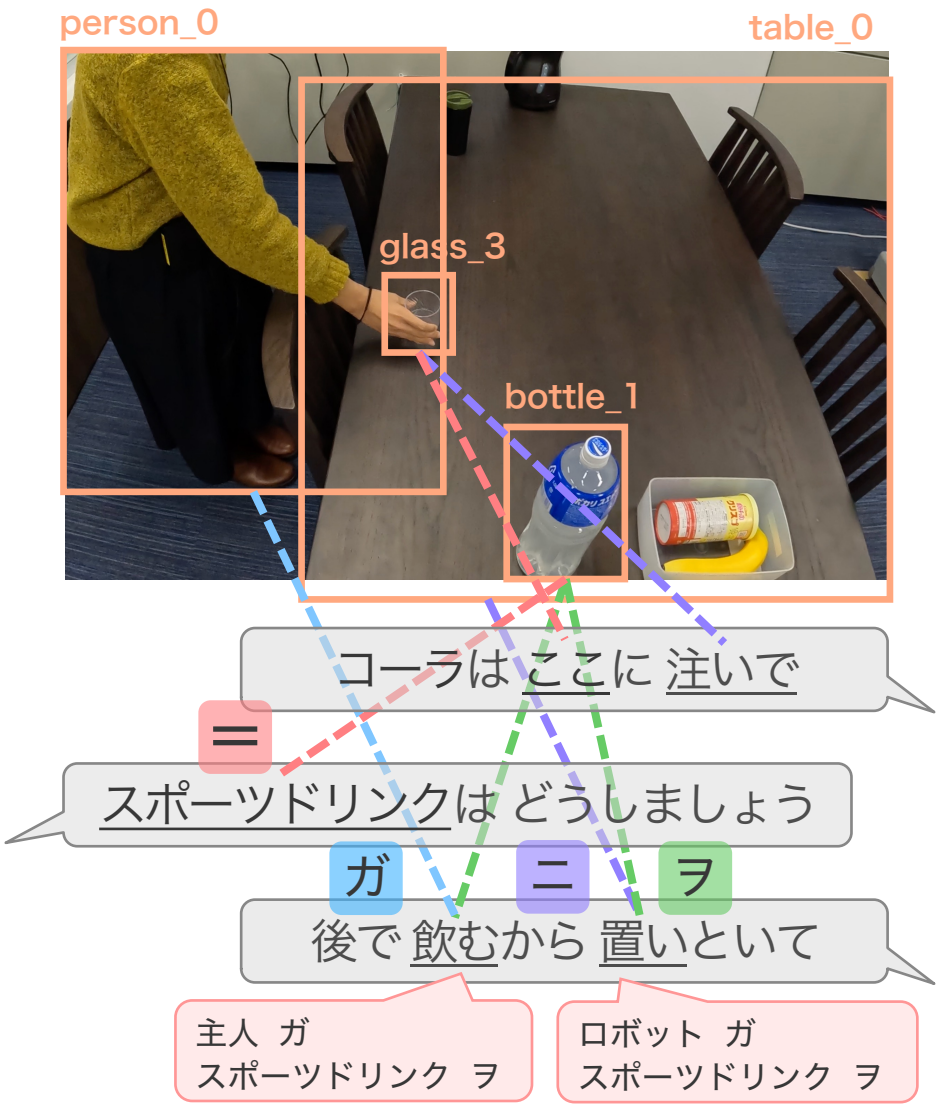
**前処理：**

- 対話音声を書き起こし
- 1人称視点動画を 1fps で画像系列に変換

**1. テキスト間照応アノテーション：**
述語項構造，橋渡し照応，共参照

**2. 物体領域アノテーション：**
物体矩形，物体クラス，インスタンスID

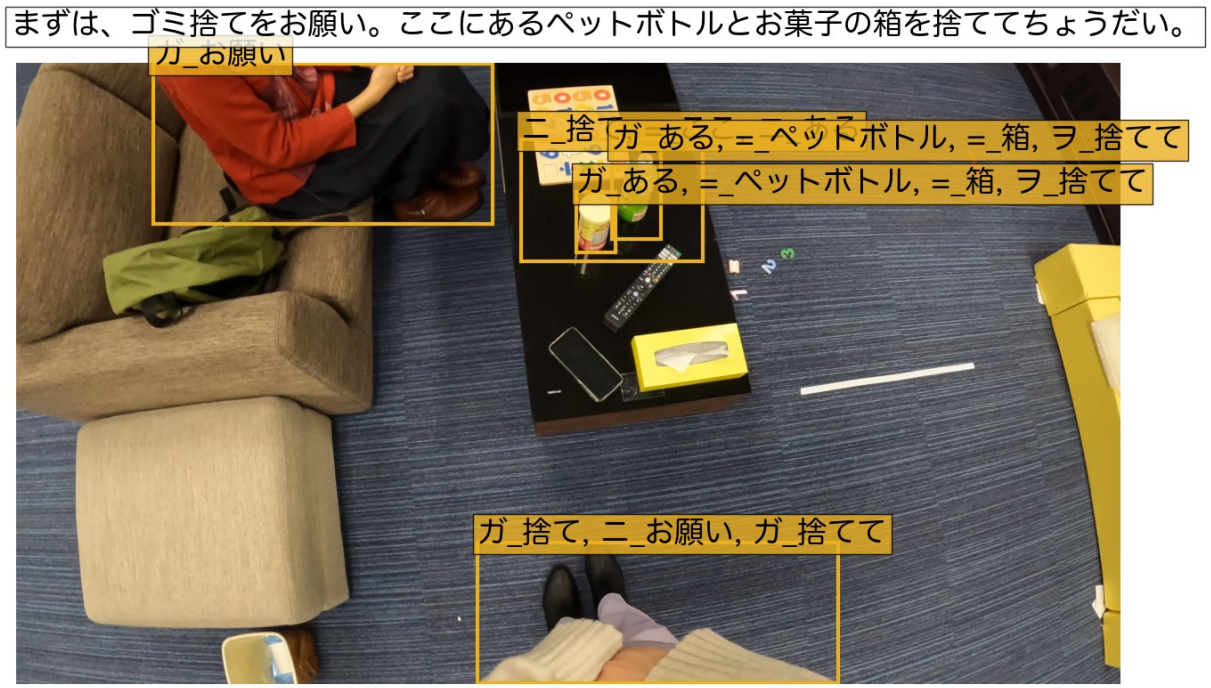
**3. テキスト・物体間参照アノテーション：**
名詞句および述語と対応する物体矩形の紐付け



# 3. テキストおよび画像アノテーション：例

発話（258発話 / 14対話）

{関係}_{名詞句 or 述語}


動画フレーム（計1,438 / 14対話）

注：実際の物体矩形には対話中の全フレーズとの関係が付与されているが,
簡単のため発話中のフレーズのみ表示している

# 4. マルチモーダル参照解析システムの構築

# マルチモーダル参照解析：タスク設定

## 入力

1人称視点画像

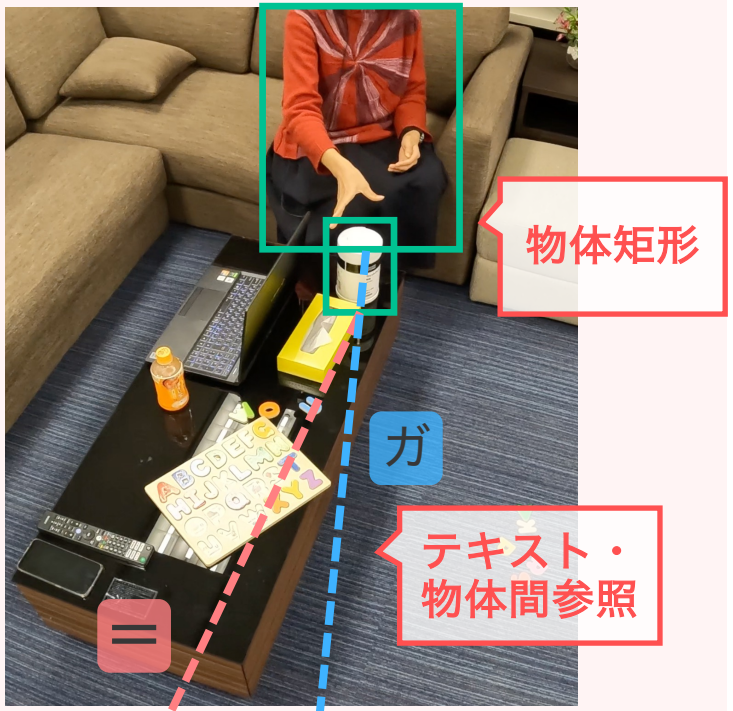
対話文脈

| 話者 | 発話 |
|---|---|
| ロボット | ： |
| 主人 | 床に落ちている物は埃をかぶっていて汚れていそうだから拭いてから戻してほしいの。 |
| ロボット | 分かりました。何で拭きましょうか？ |
| 主人 | そうね。ここにあるこれでお願いできるかしら。 |

解析対象の発話

システム

## 出力




ここにあるこれでお願いできるかしら

これ ガ
ここ ニ

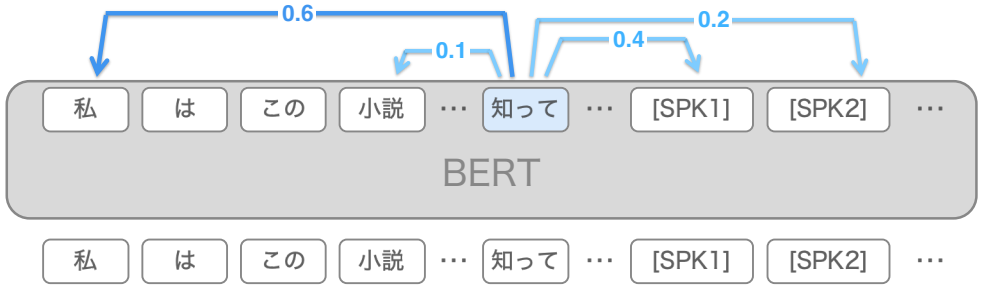
主人 ガ
ロボット ニ
拭いて ヲ

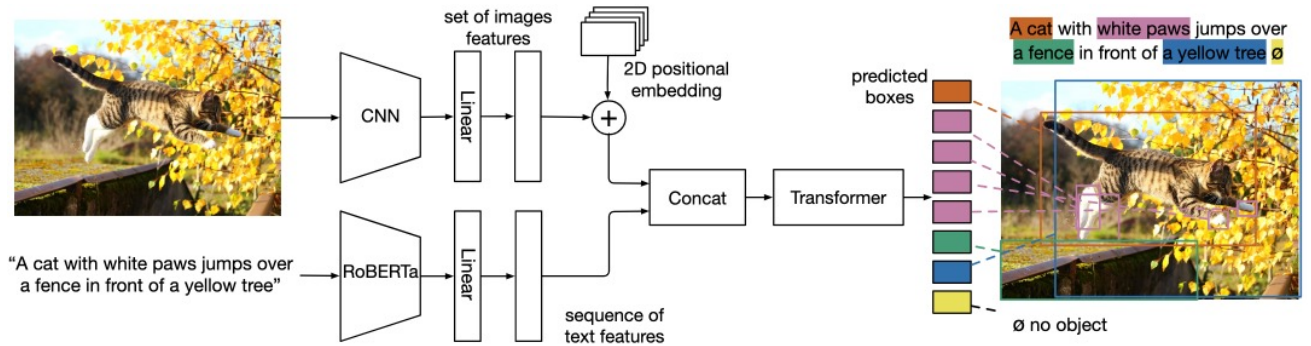
テキスト間照応

# マルチモーダル参照解析：モデル

## テキスト間照応 と テキスト・画像間参照を個別に解析

- テキスト間照応解析モデル
  - テキスト間照応データセット（KWDLC [Hangyo+12], など）で単語選択モデル [Ueda+20] を訓練
- phrase grounding モデル
  - Flickr30kEnt-JPおよび複数の英語 V&L データセットで訓練済みMDETR [Kamath+21] を fine-tuning



# マルチモーダル参照解析：実験結果

アノテーション済み14対話を用いて参照先フレーズのF値および物体矩形の再現率で評価

テキスト間照応解析モデルの結果

| タスク | 格 | 非ゼロ照応 | ゼロ照応 | 全体 |
|---|---|---|---|---|
| 述語項構造 | ガ | 0.95 | 0.36 | 0.46 |
| | ヲ | 0.86 | 0.63 | 0.73 |
| | ニ | 0.99 | 0.12 | 0.48 |
| | ガ２ | 0.86 | 0.03 | 0.06 |
| 橋渡し照応 | - | - | - | 0.60 |
| 共参照 | - | - | - | 0.64 |

phrase grounding モデルの結果

| 評価セット | Recall@1 | Recall@5 | Recall@10 |
|---|---|---|---|
| 本データセット | 0.11 | 0.17 | 0.18 |
| Flickr30kEnt-JP | 0.76 | 0.90 | 0.93 |

既存手法で本データセットを解くのは困難
- ドメインの差異のため
- 指示代名詞を含むため

ゼロ照応解析におけるF値が低い
- 対話における話者の交代が考慮されていないため

# おわりに

- 発話テキストの参照先理解のための**マルチモーダル参照解析**を提案
- アノテーション基準を策定しつつ，マルチモーダル対話データセットを部分的に構築（14対話）

今後の課題

- マルチモーダル対話データセットの拡充
- マルチモーダル参照解析システムの改善
  - phrase grounding モデルにおいてテキスト間照応解析の結果を利用
  - ゼロ照応への対応
  - 実世界でのデータ収集量の懸念を解消するためバーチャルデータも使用