E1-4

実世界における
フレーズグラウンディングモデルの評価と分析

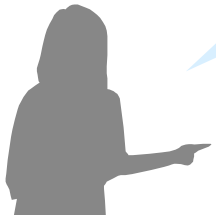
植田 暢大[1,2]　波部 英子[2]　松井 陽子[2]　湯口 彰重[3,2]　河野 誠也[2,4]
川西 康友[2,4]　黒橋 禎夫[1,2,5]　吉野 幸一郎[2,4]

[1]京都大学　[2]理化学研究所GRP　[3]東京理科大学
[4]奈良先端科学技術大学院大学　[5]国立情報学研究所

言語処理学会第30回年次大会 2024/3/12

研究背景：実世界対話における物体への参照

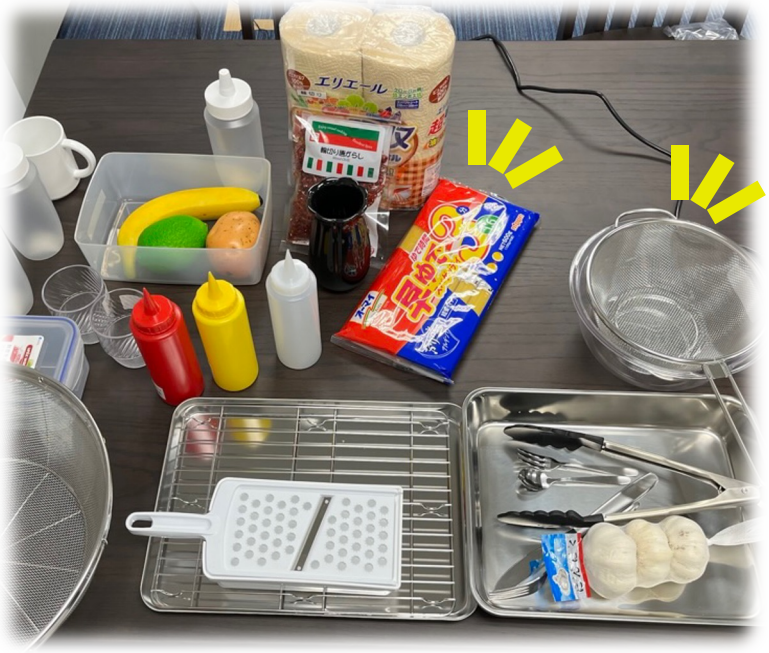
- 実世界の共同作業における対話では，**物体への参照**が頻出
- 参照先物体の特定はコンピュータによる発話理解に不可欠



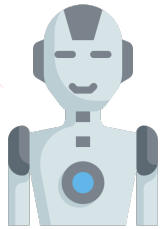
そこの**ザル**を持ってきて

ペペロンチーノね
あと**パスタ**も準備して

お願いね



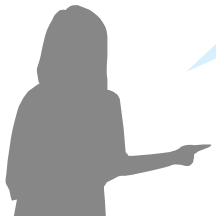
小さい方のザルですね
何を作るんですか？



分かりました
パスタの封を切っておきます

研究背景：実世界対話における物体への参照

フレーズグラウンディング：テキスト中のフレーズの参照先を対応する画像中の物体矩形として特定

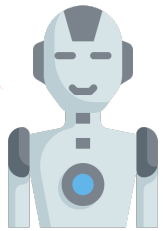


そこの**ザル**を持ってきて

ペペロンチーノね
あと**パスタ**も準備して

お願いね

小さい方のザルですね
何を作るんですか？

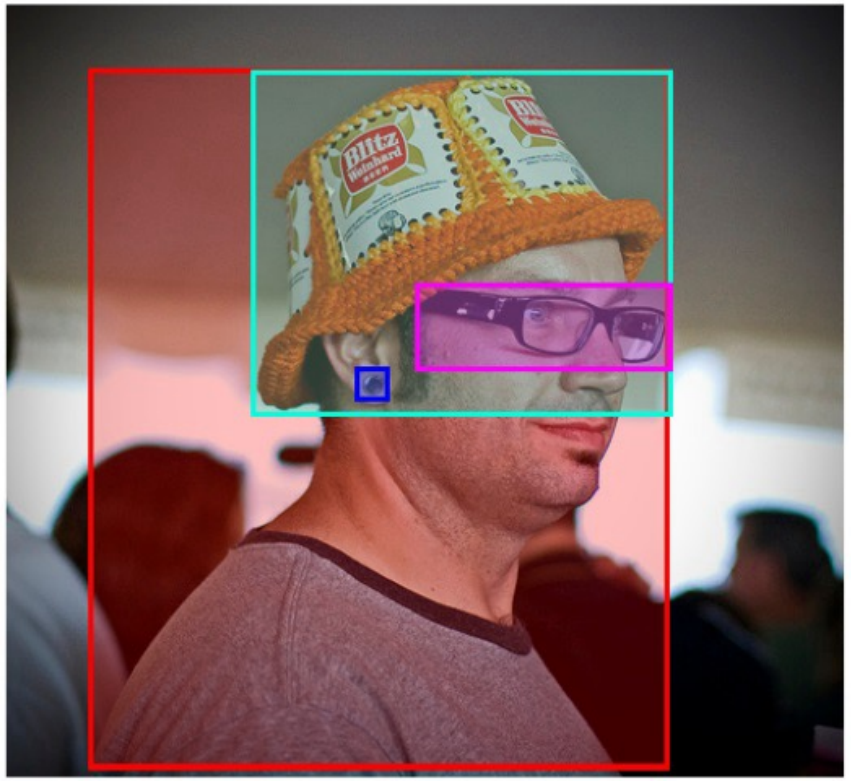


分かりました
パスタの封を切っておきます

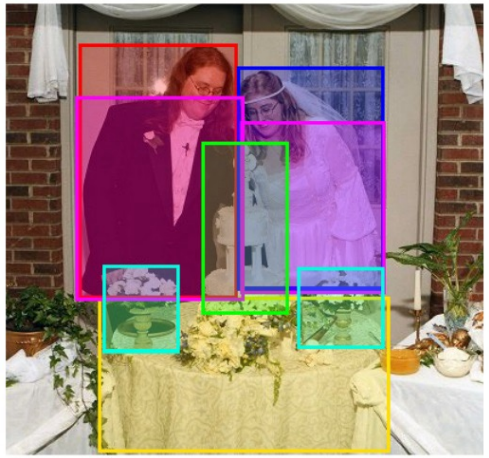
Q. 既存のモデルは実世界対話におけるフレーズグラウンディングが解けるのか？

関連研究：Flickr30k Entities データセット [Plummer+, IJCV, 2017]

画像キャプション中のフレーズに対して参照先の物体矩形を付与



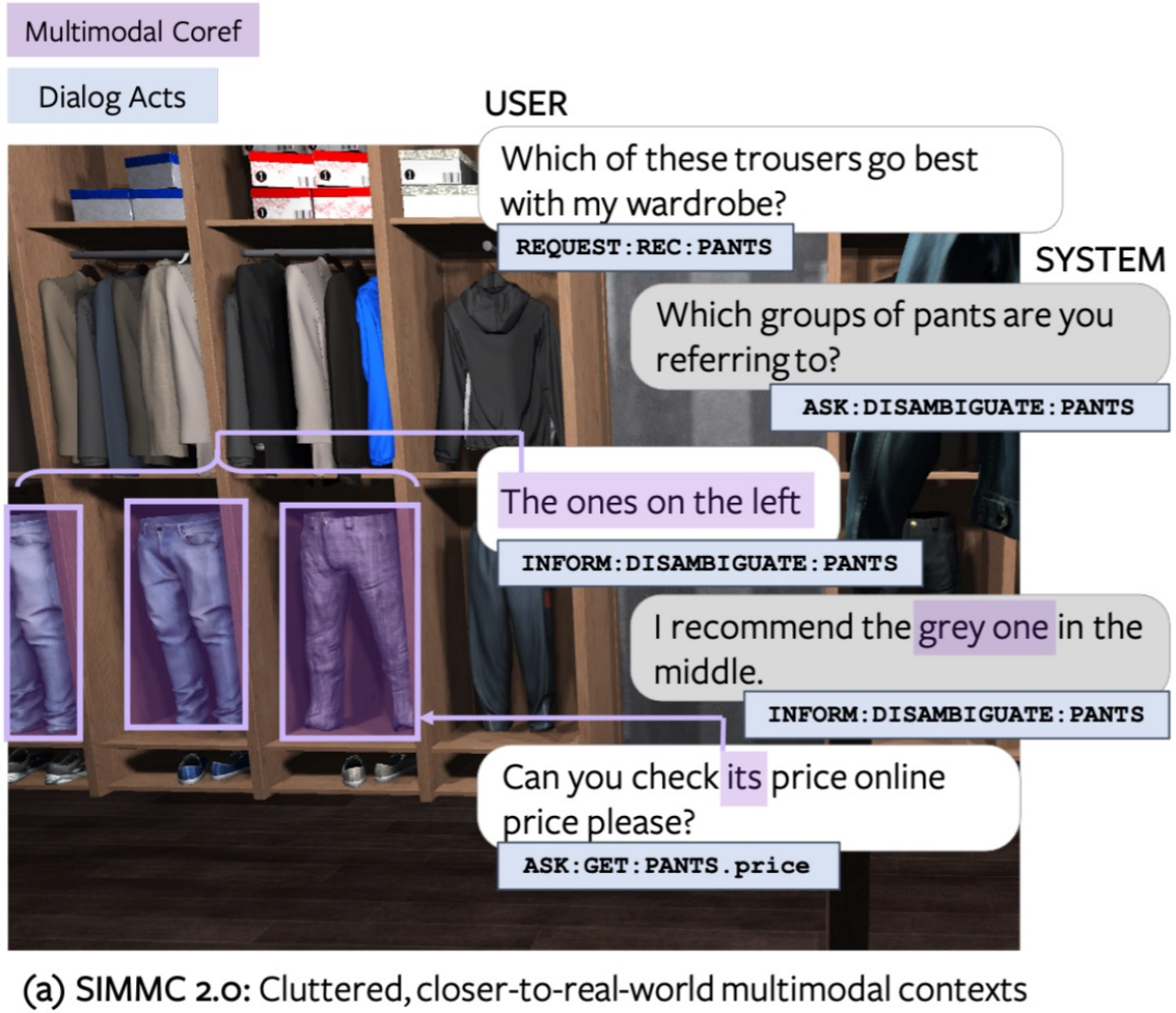
A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.



	Flickr30k Entities	実世界対話
視覚情報	（恣意的に撮影された）静止画	1人称視点動画
言語情報	キャプション	対話
既存モデルの精度	80%以上	?

関連研究：SIMMC 2.0 データセット [Kottur+, EMNLP, 2020]

買い物の場面における対話テキストとCG画像に参照関係を付与



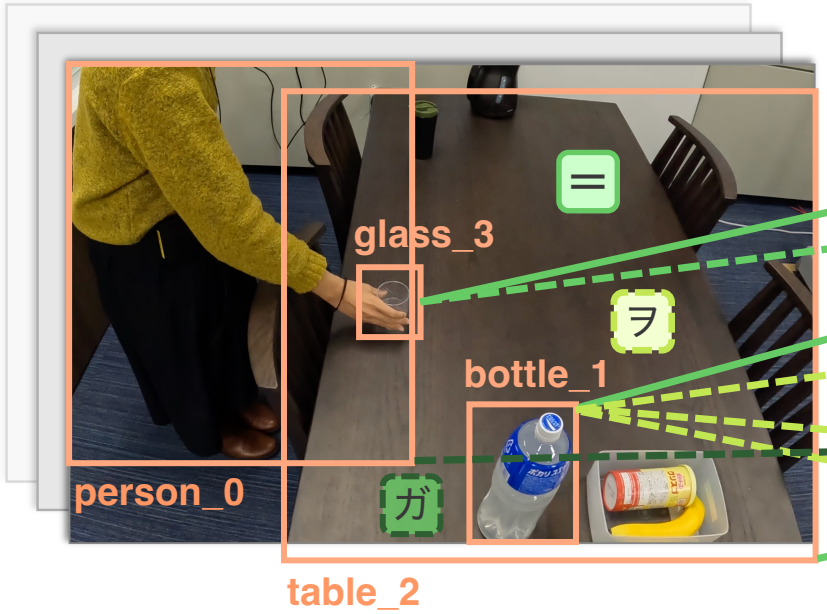
(a) SIMMC 2.0: Cluttered, closer-to-real-world multimodal contexts

	SIMMC 2.0	実世界対話
視覚情報	CGの静止画	1人称視点動画
言語情報	対話	対話
既存モデルの精度	70%以上	?

SIMMCにはCGの静止画しか含まれておらず実世界特有の視点の移動や物体の操作を含まない

J-CRe3：参照タグ付き実世界対話データセット [植田+, NLP, 2023]

- 家庭内を模した設備で主人とロボットを演じる2者に共同作業を行ってもらい，対話音声と1人称視点動画を収録
- 音声の書き起こしと1秒ごとの動画フレームにタグ付け



話者	発話書き起こし
主人	コーラは
	ここに
	注いで
ロボット	スポーツドリンクは
	どうしましょう
主人	後で
	飲むから
	置いといて

ガ格	ヲ格	ニ格	共参照
ロボット	コーラ	ここ	
ロボット	スポーツドリンク		
主人	スポーツドリンク		
ロボット	スポーツドリンク		

物体矩形および
インスタンスID

テキスト・物体間参照

テキスト間照応

タスク設定：J-CRe3におけるフレーズグラウンディング

解析対象
フレーム

Frame #41

Frame #42

Frame #43

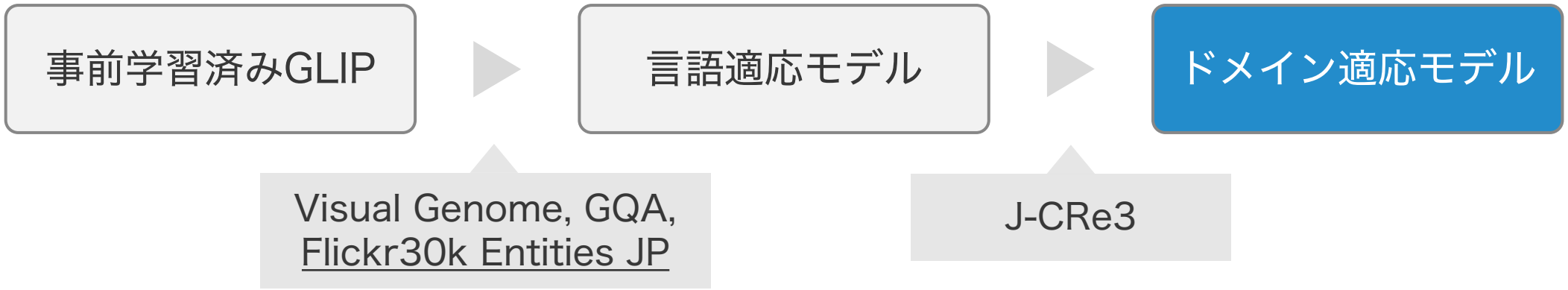
話者	発話 開始時刻	発話
ロボット		⋮
主人	0:24:961	床に落ちている物は埃をかぶっていて汚れていそうだから拭いてから戻してほしいの。
ロボット	0:34:649	分かりました。何で拭きましょうか？
主人	0:40:077	そうね。ここにあるこれでお願いできるかしら。
ロボット	0:43:366	分かりました。
主人		⋮

入力コンテキスト

解析対象発話

ベースライン：フレーズグラウンディングモデル GLIP の fine-tuning

- GLIP [Li et al., CVPR, 2022] は物体検出をフレーズグラウンディングと同一の枠組みで扱い大量の画像・テキストペアで訓練されたモデル
- GLIP を fine-tuning し，画像・テキストペアごとに独立に解析



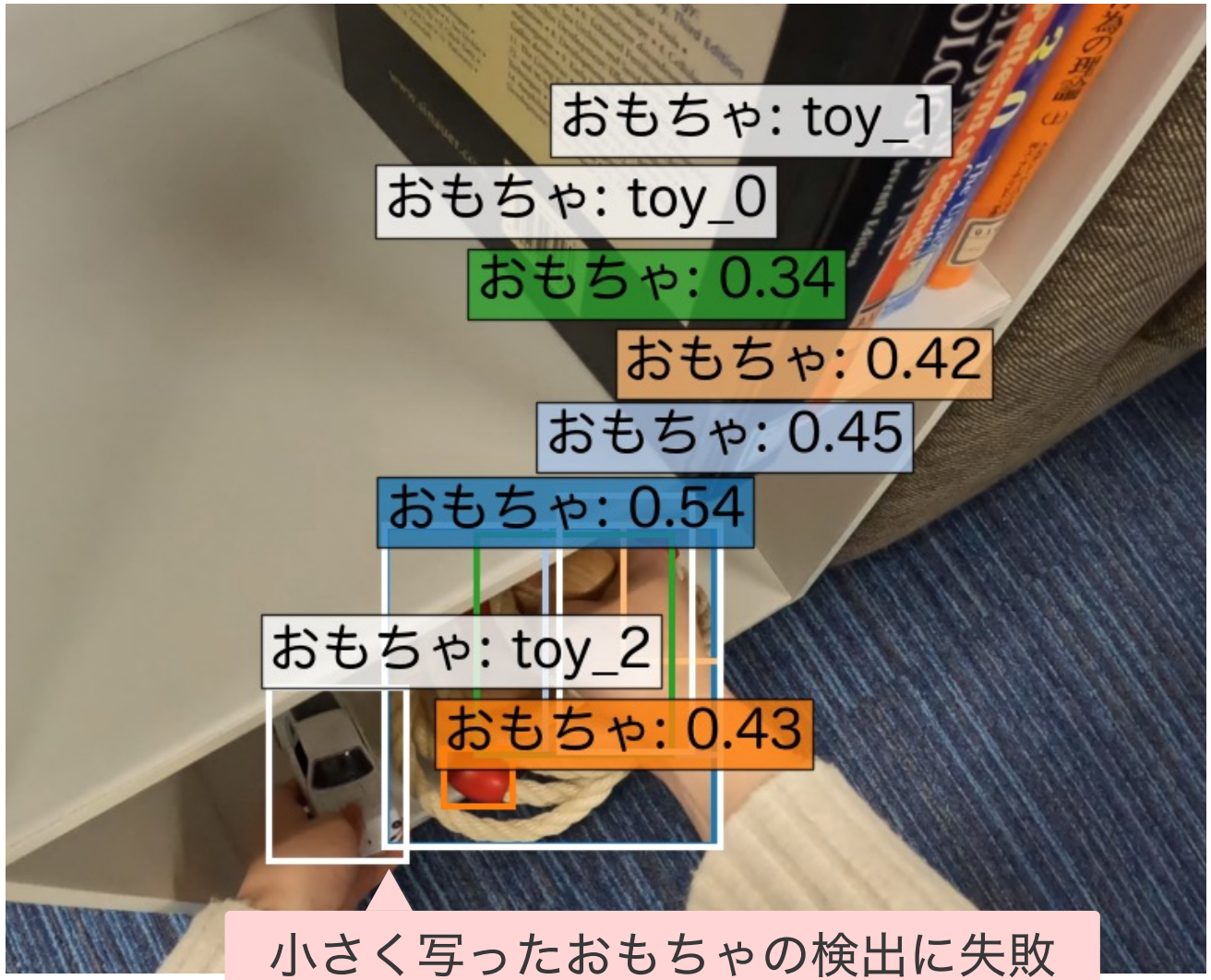
評価セット	Recall@1	Recall@5	Recall@10
Flickr30k Entities JP [Nakayama+20]	0.695	0.881	0.915
J-CRe3	0.477	0.700	0.764

標準的なベンチマークと比べスコアが大幅に低い

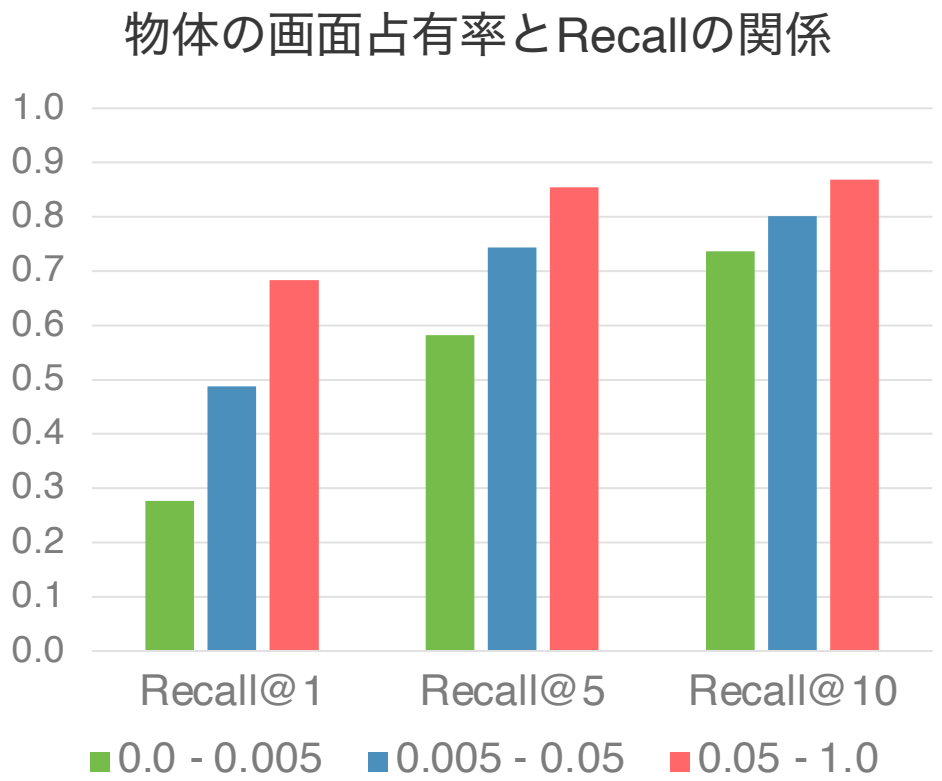
ドメイン適応モデルが予測した物体矩形のRecall

エラー分析（1/2）：小さく写っている物体における精度

「あそうそう、**おもちゃ**を置いた後で二段目の本を取ってきてほしいの」



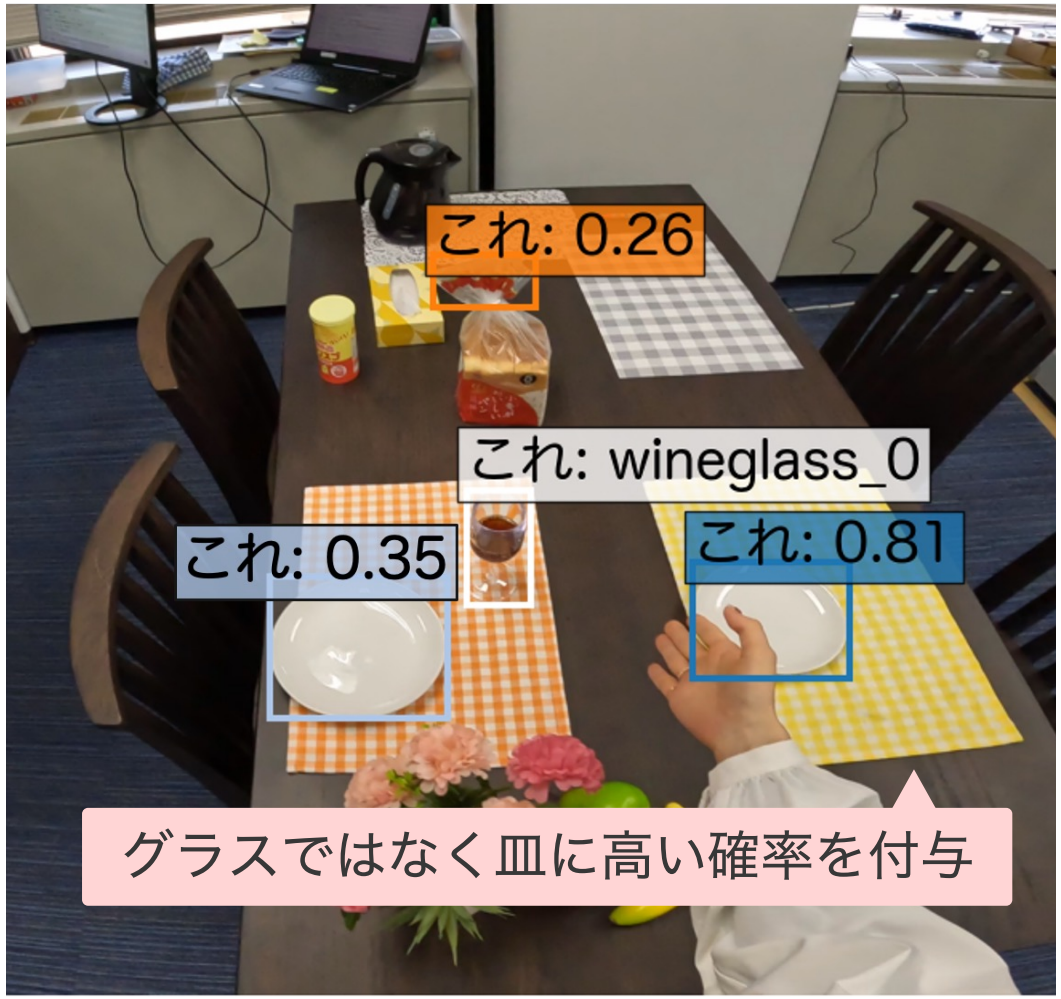
小さく写ったおもちゃの検出に失敗

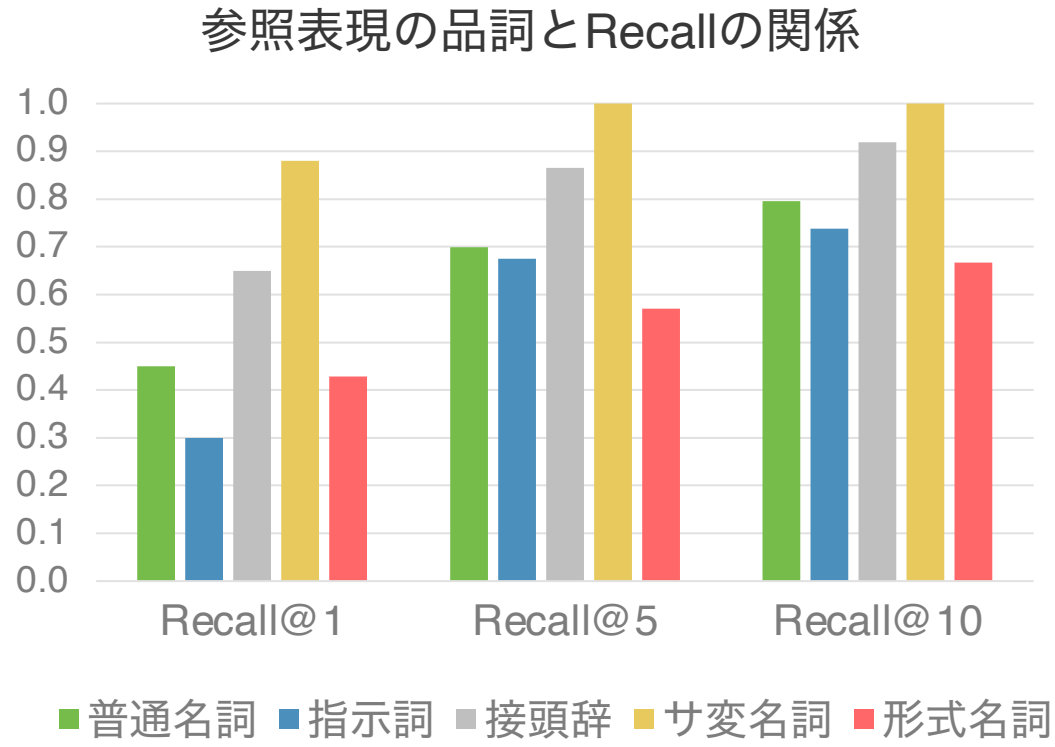


小さく写る物体に対する精度が低い

エラー分析（2/2）：曖昧な表現に対する精度

「**これ**はまだ飲みますか？洗っちゃいますか？」





指示詞や形式名詞などの曖昧な表現に対する精度が低い

提案手法（1/2）：インスタンスIDの利用

複数物体追跡器から得られるインスタンスIDを用いてシステム出力を補正

Frame #42

インスタンス ID: paper_towel_3

Frame #43

0.9

0.7 → 0.9

話者	発話
主人	⋮
ロボット	このウェットティッシュでお願いできるかしら。
主人	そうね。ここにあるこれでお願いできるかしら。
ロボット	⋮

提案手法（2/2）：共参照クラスタの利用

共参照解析器から得られる共参照クラスタを用いてシステム出力を拡張

Frame #42

Frame #43

0.9

0.7 → 0.9

話者	発話
ロボット	：
主人	ウェットティッシュがあるから、 これでお願いできるかしら。 共参照
ロボット	分かりました。 これで拭けば良いんですか？
主人	：

実験結果（1/2）

複数物体追跡（MOT）および共参照（Coref.）を考慮した場合のフレーズグラウンディングの結果

手法	テストセット（9 対話）			開発セット（6 対話）		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
ベースライン	0.477 (194/407)	0.700 (285)	0.764 (311)	0.416 (144/346)	0.711 (246)	0.824 (285)
+ MOT (gold)	0.477 (194/407)	0.742 (302)	0.813 (331)	0.457 (158/346)	0.754 (261)	0.876 (303)
+ Coref. (gold)	0.474 (193/407)	0.700 (285)	0.764 (311)	0.416 (144/346)	0.697 (241)	0.783 (271)
+ MOT (gold) + Coref. (gold)	0.479 (195/407)	0.759 (309)	0.826 (336)	0.465 (161/346)	0.832 (288)	0.954 (330)
+ MOT	0.474 (193/407)	0.700 (285)	0.764 (311)	0.405 (140/346)	0.711 (246)	0.824 (285)
+ Coref.	0.474 (193/407)	0.700 (285)	0.764 (311)	0.416 (144/346)	0.702 (243)	0.812 (281)
+ MOT + Coref.	0.472 (192/407)	0.700 (285)	0.764 (311)	0.405 (140/346)	0.702 (243)	0.812 (281)

解析器の出力を使用した場合のゲインはなし

実験結果（2/2）

共参照（Coref.）および複数物体追跡（MOT）を考慮した場合のフレーズグラウンディングの結果

手法	テストセット（9 対話） Recall@1	Recall@5	Recall@10	開発セット（6 対話） Recall@1	Recall@5	Recall@10
ベースライン	0.477 (194/407)	0.700 (285)	0.764 (311)	0.416 (144/346)	0.711 (246)	0.824 (285)
+ MOT (gold)	0.477 (194/407)	0.742 (302)	0.813 (331)	0.457 (158/346)	0.754 (261)	0.876 (303)
+ Coref. (gold)	0.474 (193/407)	0.700 (285)	0.764 (311)	0.416 (144/346)	0.697 (241)	0.783 (271)
+ MOT (gold) + Coref. (gold)	0.479 (195/407)	0.759 (309)	0.826 (336)	0.465 (161/346)	0.832 (288)	0.954 (330)
+ MOT	0.474 (193/407)	0.700 (285)	0.764 (311)	0.405 (140/346)	0.711 (246)	0.824 (285)
+ Coref.	0.474 (193/407)	0.700 (285)	0.764 (311)	0.416 (144/346)	0.702 (243)	0.812 (281)
+ MOT + Coref.	0.472 (192/407)	0.700 (285)	0.764 (311)	0.405 (140/346)	0.702 (243)	0.812 (281)

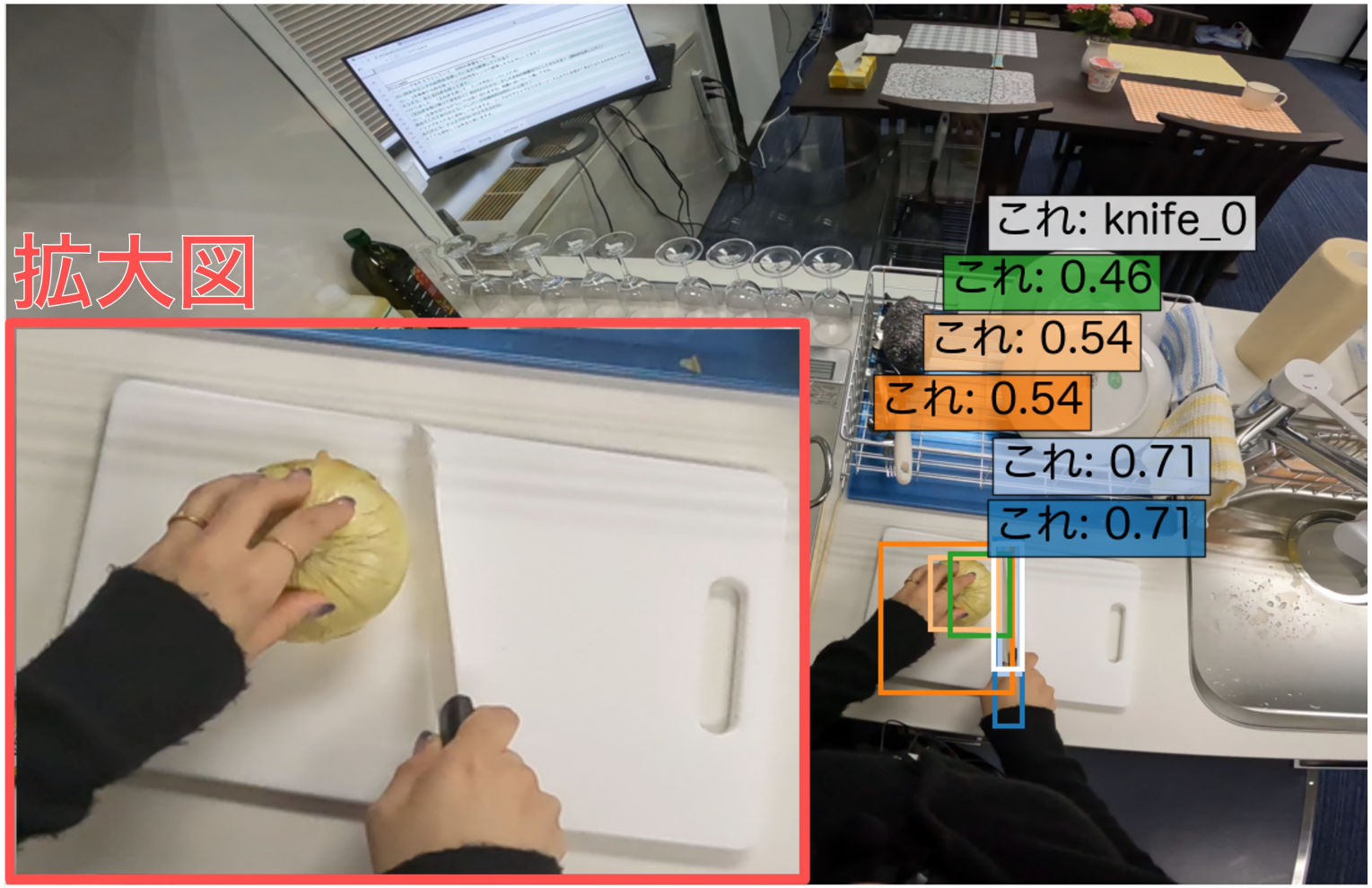
goldデータを使用した場合はゲインあり

今後解析器の性能向上によって gold を用いない設定でも性能向上の可能性を示唆

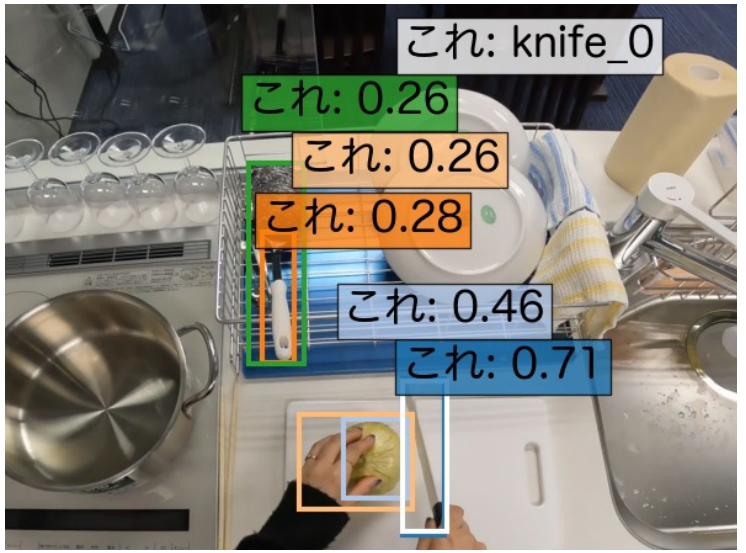
事例分析：gold を用いることで解析に成功した例

「**これ**はよく切れますね。綺麗に研いでいて偉いですね。」

Frame #51



Frame #50



包丁が正しくグラウンディングされた前フレーム (↑) の情報から，包丁の確率が 0.45 -> 0.71 に改善

おわりに

- 参照タグ付き実世界対話データセット（J-CRe3）を利用して既存のフレーズグラウンディングモデルの性能を評価・分析
- 分析に基づき，共参照および複数物体追跡を考慮した手法を提案

今後の課題

- 複数物体追跡器などの解析モジュールの性能向上
- 述語項構造など共参照以外のフレーズ間関係の利用
  - 格フレームから得られる選択選好知識など
- フレーズグラウンディングの結果を用いた共参照解析や複数物体追跡の改善


https://github.com/riken-grp/J-CRe3