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 大学院情報学研究科 [2] 理化学研究所 ガーディアンロボットプロジェクト
[3] 奈良先端科学技術大学院大学 情報科学領域
{ueda,kuro}@nlp.ist.i.kyoto-u.ac.jp
{hideko.habe,akishige.yuguchi,seiya.kawano}@riken.jp
{yasutomo.kawanishi,koichiro.yoshino}@riken.jp

概要

実世界で話者同士が視覚情報を共有しながら行う対話では，物体に対する参照表現が多く出現する．このような表現の参照先の解決のため，本研究ではマルチモーダル参照解析タスクを提案し，本タスクのためのマルチモーダル対話データセットを構築する．本データセットは実世界における対話動画および音声を基にしており，対話テキスト中のフレーズと一人称視点動画におけるフレーム内の物体領域が紐付けられている．この紐付けには同一のものを指す関係だけでなく，述語と項の関係や橋渡し照応関係も含まれる．

1 はじめに

実世界で人間と対話を通して協働するロボットの実現には，視覚的文脈を含めた人間の発話理解が不可欠である．例えば「そのコップを取って」という発話を例にとると，「コップ」のテキスト上の意味を理解するだけでは不十分であり，実世界において参照しているコップの実体を知ることが必須となる．このように，発話中の表現と参照関係を持っている実世界における物体を特定することは発話理解の基礎となる．

発話中の表現と視覚情報を紐付けた対話データセットとして SIMMC 2.0 [1] がある．SIMMC 2.0 は 1,566 枚の CG 画像と 11,244 の対話テキストに対してテキスト中の名詞句と参照先の画像中の物体を囲む矩形が紐付けられたデータセットである．規模は大きいものの，SIMMC 2.0 は話者の写っていない静止画に基づいており，物体の移動や操作が視覚的に提示されていない．また，SIMMC 2.0 の参照関係は

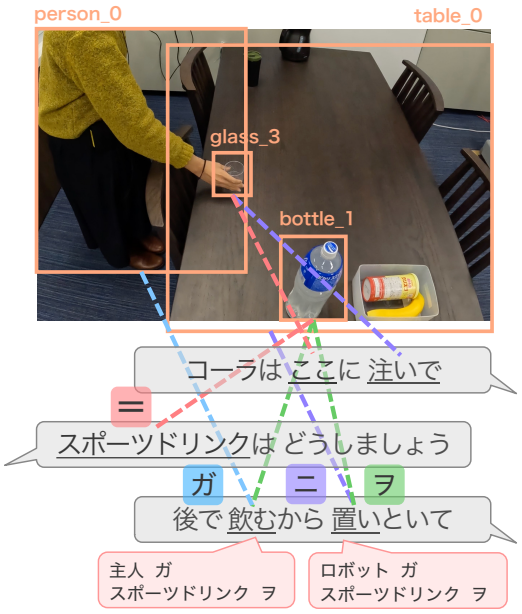

図 1 マルチモーダル対話データセットの例．物体領域（橙色矩形），テキスト間照応関係（桃色吹き出し），テキスト・物体間参照関係（破線）が付与されている．物体領域にはクラス名とインスタンス ID が付与されている．

テキストとして表出している名詞句にのみ付与されている．一方，日本語ではゼロ照応とよばれる参照元の表現が省略される現象が頻出する．例えば「そのコップを持ってきて」という発話では，持ってくる目的地にあたるニ格の項が省略されている．このような省略された表現を SIMMC 2.0 は対象としていない．

本研究では，実世界での物体操作を伴う対話においてゼロ照応も総合的に扱うマルチモーダル参照解析を提案し，そのためのデータセットを構築する．データセット構築のため，まずクラウドソーシングで対話シナリオを収集する．続いて，収集したシナリオに沿って実世界で対話音声と対話動画を収録

する．最後に，対話音声の書き起こしと対話動画のフレームに参照関係を付与する．具体的には図 1 の「スポーツドリンク」のように，対話テキスト中の名詞句に対して参照している物体を囲む矩形（物体矩形）を紐付ける．加えて本研究ではゼロ照応も扱うため，テキスト中に言及がない場合でもテキスト中の述語からそのガ格やヲ格に対応する物体を紐付ける．

データセットの評価のため，既存手法 [2] を用いて名詞句のグラウンディングタスクを解いた．結果，物体検出の再現率は 0.18 程度であり，本データセットが非常に挑戦的であることが示された．現在，収録済みデータは 24 対話，フルアノテーション済みデータは 7 対話にとどまるが，今後さらに拡張させる予定である．

2　マルチモーダル参照解析

本研究では，実世界での対話的協働を想定したマルチモーダル参照解析タスクを提案する．本タスクでは，動画あるいは画像と，対応するテキストが入力として与えられる．そのとき，テキスト中の名詞や述語における参照関係を，参照先として動画や画像中の物体も含めて解析する．本タスクは，テキスト間照応解析，物体検出，テキスト・物体間参照解析の 3 つのサブタスクから構成される．

テキスト間照応解析は，テキスト中の単語や句の間に存在する照応・共参照関係を解析するタスクである．本研究では先行研究 [3, 4] にならい，述語項構造，共参照，橋渡し照応関係を解析する．述語項構造は述語を中心とし，その述語の「誰が」や「何を」に相当する項からなる関係である．図 1 では「飲む」と「置い」について述語項構造が示されている．共参照関係は実世界において同一の実体を指し示す名詞間の関係である．橋渡し照応関係はある名詞（照応詞）と，その必須的な意味を補完する異なる名詞（先行詞）との関係である．

物体検出は，画像中から参照されている物体が存在する領域を特定するタスクである．図 1 においては画像中の物体矩形を推定することに対応する．入力が動画の場合は動画中のそれぞれのフレームに対して同様の処理を行う．

テキスト・物体間参照解析は，テキスト間照応解析における照応・共参照の対象を物体検出によって特定された物体領域から選択するタスクである．図 1 においては単語と物体矩形を結ぶエッジを推定することに対応する．表現が直接参照している対象を画像中から検出するタスクは phrase grounding [2, 5] や referring expression comprehension [6] として知られるが，本研究では述語項構造や橋渡し照応など，ゼロ照応を含む間接的な関係も扱う．

3　マルチモーダル対話データセット

本研究では，マルチモーダル参照解析のためのマルチモーダル対話データセットを構築する．本データセットは，実世界における 2 者の対話シーンにおいて動画，音声を収録し，音声書き起こし，照応・参照関係アノテーションを付与したデータセットである．対話内容は，人間とお手伝いロボットの対話を想定する．対話場面は，家庭内のリビングとダイニングを模した 2 種類である．本節ではマルチモーダル対話データセットの構築方法について順に述べる．

3.1　対話シナリオ収集

多様かつ現実的な対話シナリオを得るため，クラウドソーシングを利用してシナリオを収集した．クラウドソーシングタスクではワーカーに対話収録に使用する部屋の状況と使用可能な物体の写真を提示した．その上で，人間とロボットの発話およびその際の動作や場面状況を記述してもらった．発話数は長すぎず，かつ対話が十分な文脈を持つよう 10–16 発話に制限した．収集したシナリオを実行可能性，十分な頻度の参照表現，十分な粒度の場面状況説明，の 3 つの観点からフィルタリングし，残ったシナリオを自然な対話になるよう修正した．付録 A に修正後のシナリオの例を示す．

3.2　対話収録

収集したシナリオに基づいて実際に対話を行い，データを収録した．シナリオは人対ロボットを想定しているが，今回はロボット役も人間の演者に依頼した．演者にはできる限りシナリオを暗記してもらい，対話中の振る舞いが自然になるようにした．なお，データセットにおいては実際に行われた発話に対応するよう，アノテーションの際にシナリオを元に台詞を修正する．

収録はリビングとダイニングを模した設備が備え付けられた実験室で行った．2 人の演者にはそれぞれピンマイクを付けてもらい，発話を録音した．ロボット役の演者には頭部にカメラを付けてもらい，

対話中の 1 人称視点動画を撮影した．さらに，実験室に定点カメラを 4 箇所設置し，部屋全体の様子を撮影した．

3.3 アノテーション

収録した対話音声と動画に対してマルチモーダル参照解析のためのアノテーションを行った．まず対話音声はテキストに書き起こし，1 人称視点動画は 1 秒ごとにフレームを抽出し画像系列に変換した．以下では，テキスト間照応解析，物体検出，テキスト・物体間参照解析に対応するアノテーションをそれぞれテキスト間照応アノテーション，物体領域アノテーション, テキスト・物体間参照アノテーションとよぶ．

テキスト間照応アノテーション　書き起こされた対話テキストに対して述語項構造・共参照・橋渡し照応関係を付与した．これらはテキストの結束性を構成するエンティティ間の重要な関係である [4]．アノテーションの基準は京都大学ウェブ文書リードコーパス [7, 8] に準拠した．

物体領域アノテーション　動画から抽出された全フレームについて，物体に物体矩形を付与した．また，それぞれの物体矩形に対して物体のクラス名およびインスタンス ID を付与した．クラス名は物体認識タスクにおいて広く利用される LVIS データセット [9] に定義されている 1,203 のクラスを使用した．アノテーションコストを減らすため，付与対象の物体はゼロ照応も含め対話中で参照された物体に限定した．また，一般物体認識器 Detic [10] の学習済みモデル[1)] と複数物体追跡器 SORT [11] を使用してシルバーアノテーションを事前付与し，手作業で修正した．

テキスト・物体間参照アノテーション　テキスト中の (1) 名詞句および (2) 述語と，画像中の物体矩形のすべての組み合わせについて参照関係を付与した．(1) 名詞句については，直接参照している物体および橋渡し照応関係にある物体に矩形を付与した．(2) 述語についてはその項に対応する物体矩形を格ごとに付与した．

テキスト・物体間参照アノテーションは付与対象の関係が非常に多くなる．しかし，付与済みのテキスト間照応・共参照関係とインスタンス ID を利用することで大部分のアノテーションを省くことができる．例えば，ある動画フレーム中の「コップ」に対して参照関係を付与した場合を考える．このとき，別フレームに同じインスタンス ID を持つ「コップ」が出現したとしても自動的に参照関係を付与できる．

実世界対話における特有の現象として「ここ」や「あそこ」など，場所を指す指示代名詞の使用が挙げられる．指し示された場所の特定は，人間との協働において不可欠である．本研究では，このような名詞を領域参照表現とよび，テキスト・物体間参照アノテーションにおいて対応する領域を付与する．この領域は物体領域とは異なり矩形が一意に定まらないため，その位置・大きさはアノテータの主観に依存する．そのため，特殊なクラス名である「region」を付与し，特殊な物体矩形として扱い，他の物体矩形とは区別する．

表 1, 2, 3 にそれぞれアノテーション済みの対話における統計情報を示す．テキスト・画像の両モダリティにおいて十分な参照関係が存在することが分かる．付録 A にデータセットに含まれる対話数を示す．

表 1　言語アノテーションの統計情報（24 対話）．

	最小	平均	最大
述語数	17.0	33.5	74.0
項の数	39.0	72.8	129.0
共参照名詞数	4.0	14.3	29.0
橋渡し照応詞数	1.0	5.2	15.0

表 2　画像アノテーションの統計情報（12 対話）．

	最小	平均	最大
フレーム数	70.0	100.1	142.0
物体矩形数	205.0	541.3	1042.0
フレームあたりの物体矩形数	2.7	5.5	8.7
オブジェクトインスタンス数	5.0	13.9	26.0

表 3　物体矩形と名詞句・述語の関係数の統計（7 対話）．「=」は直接の参照関係を表す．

	最小	平均	最大
=	1.0	17.0	40.0
ガ格	10.0	24.7	41.0
ヲ格	2.0	16.1	53.0
ニ格	6.0	11.1	19.0
ガ 2 格	4.0	6.6	12.0
デ格	0.0	0.6	2.0
ト格	0.0	0.1	1.0
橋渡し	0.0	3.7	9.0

1) https://github.com/facebookresearch/Detic/blob/main/docs/MODEL_ZOO.md

4 実験

物体検出とテキスト・物体間参照解析は phrase grounding をサブタスクとして含む．本節では既存の phrase grounding モデルを使用した本タスクの難しさを評価する実験について述べる．述語項構造や橋渡し照応に対応する間接的な関係はモデルの性質上扱わない．

4.1 タスク設定

phrase grounding は，テキストと画像が与えられたときテキスト中のフレーズに対応する画像中の物体矩形を推定するタスクである [2, 5]．マルチモーダル対話データセットは画像ではなく動画を元にした画像系列から構成される．すなわち，1 フレーズについてグラウンディング対象の画像が複数存在する．本実験では，簡単のためフレーズが含まれる発話区間内の画像に限定する．

4.2 実験設定

phrase grounding モデルとして MDETR [2] を使用した．MDETR は Transformer アーキテクチャ [12] をベースとした，物体検出と phrase grounding を end-to-end で行うモデルである．MDETR は学習済みモデル[2)]が公開されているが，英語テキストで学習されたモデルであるため直接利用できない．そこで我々は， Flickr30k Entities JP データセット [13] を使用してこのモデルを fine-tuning した．このデータセットは phrase grounding において標準的に使用される Flickr30k Entities データセット [14] を日本語に翻訳したものである．なお，マルチモーダル対話データセットは規模が小さいため fine-tuning には使用しなかった．

評価指標は Recall@k を使用した．MDETR は，それぞれのフレーズについて複数の物体矩形とその予測確率を出力する．Recall@k は正解の物体矩形のうち，出力された予測確率上位 k 件の物体矩形に含まれるものの割合である．ここで，先行研究 [2] にならい予測された物体矩形が正解の物体矩形と 0.5 以上の Intersection-over-Union (IoU) を持つ場合に両者が一致すると判断した．

表 4 phrase grounding モデルの精度．

評価セット	R@1	R@5	R@10
Flickr30k Entities JP	0.76	0.90	0.93
本データセット（7 対話）	0.18	0.29	0.33

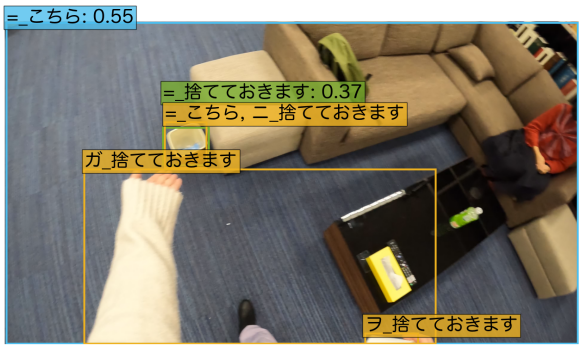

図 2 「こちらに捨ててておきますね。」という発話に対する phrase grounding モデルの解析例．黄色の物体矩形が正解データを表し，それ以外はシステムの出力である．システム出力にはシステムの予測確率が記載されている．

4.3 実験結果

実験結果を表 4 に示す．Flickr30k Entities JP で評価した場合に比べ，マルチモーダル対話データセットにおける Recall@k のスコアは大幅に低い．このことから，既存手法では本タスクを解くことが難しいことが分かる．

図 2 にマルチモーダル対話データセットにおける解析例を示す．「コップ」や「ペットボトル」など具体的な物体名については正しく解析できた事例が多かったが，図に示した「こちら」等の曖昧な表現については正しい事例が少なかった．これは Flickr30k Entities データセットに出現する物体名の多くが具体的であるためと考えられる．

5 おわりに

本研究では，実世界において人間と対話しつつ協働するロボットの実現を目指し，テキスト・画像のマルチモーダル参照解析タスクを提案した．さらに，本タスクを解くためのマルチモーダル対話データセットを構築した．本データセットは実世界における実際の対話を基に作成されており，より実用的な発話理解システムの実現に役立つと期待される．今後は，データセットの拡張と並行して，本タスクのための解析モデルに取り組む．

2) https://github.com/ashkamath/mdetr#pre-training

謝辞

本研究は，京都大学科学技術イノベーション創出フェローシップ事業の助成を受けたものである．本研究の一部は JSPS 科研費 22H03654 の支援を受けたものである．

参考文献

[1] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 4903–4912, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 1780–1790, 2021.

[3] Masato Umakoshi, Yugo Murawaki, and Sadao Kurohashi. Japanese zero anaphora resolution can benefit from parallel texts through neural transfer learning. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 1920–1934, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[4] Nobuhiro Ueda, Daisuke Kawahara, and Sadao Kurohashi. BERT-based cohesion analysis of Japanese texts. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 1323–1333, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[5] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, **Computer Vision – ECCV 2020**, pp. 752–768, Cham, 2020. Springer International Publishing.

[6] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. **IEEE Transactions on Multimedia**, Vol. 23, pp. 4426–4440, 2020.

[7] 萩行正嗣, 河原大輔, 黒橋禎夫. 多様な文書の書き始めに対する意味関係タグ付きコーパスの構築とその分析. 自然言語処理, Vol. 21, No. 2, pp. 213–248, 2014.4.

[8] Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. Building a diverse document leads corpus annotated with semantic relations. In **Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (PACLIC)**, pp. 535–544, 2012.

[9] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, 2019.

[10] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, **Computer Vision – ECCV 2022**, pp. 350–368, Cham, 2022. Springer Nature Switzerland.

[11] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In **2016 IEEE International Conference on Image Processing (ICIP)**, pp. 3464–3468, 2016.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **NIPS2017**, pp. 5998–6008. 2017.

[13] Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. A visually-grounded parallel corpus with phrase-to-region linking. In **Proceedings of The 12th Language Resources and Evaluation Conference**, pp. 4204–4210, Marseille, France, May 2020. European Language Resources Association.

[14] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. **IJCV**, Vol. 123, No. 1, pp. 74–93, 2017.

付録

表 5　収集したシナリオの例．括弧内は場面状況であり台詞としては使用しない．

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。

表 6　データセットに含まれる対話数．

設定	全シナリオ	収録済み	アノテーション済み
リビング	30	12	3
ダイニング	30	12	4

A　データセットの詳細

表 5 に収集した対話シナリオの例を示す．括弧内のテキストは場面状況を表す．括弧外の発話には「あそこ」や「それ」などの参照表現が含まれ，視覚情報も含めなければ理解が困難な対話になっている．

表 6 に現在のデータセットに含まれる対話数を示す．将来的にはシナリオ数も含め，さらに拡張する予定である．