実世界対話における参照関係の統合的解析

稲積 駿 [1,2]　植田 暢大 [3,*]　吉野 幸一郎 [4,2,1]
[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
[3] 京都大学　[4] 東京科学大学
inazumi.shun.in6@naist.ac.jp　ueda@nlp.ist.i.kyoto-u.ac.jp
koichiro.yoshino@riken.jp

概要

本研究では，実世界における日本語対話の曖昧性解消を目的として，マルチモーダル参照解析の性能向上に寄与する要素を検討し，またモデルに組み込む．具体的には，メンション間およびメンション・物体間の参照関係を統合的に解析するフレームワークを提案する．実験では，共参照解析や照応解析といったメンション間の学習がメンション・物体間の解析に与える利得を明らかにした．

1　はじめに

対話を介して実世界でユーザと協働が可能な対話システムに向けて，対話テキストに含まれる参照表現 (メンション) とメンションが参照する物体を特定する試みがなされている [1, 2, 3, 4]．マルチモーダル参照解析は，メンション同士の参照関係に加えメンションから物体への参照関係[1]を特定するタスクである [1]．例えば，図 1 の状況においてシステムは，自身の観測から “そのコップ” や “取って” が参照する他のメンションや物体を特定する．これにより，対話に含まれる “何を誰が誰にどうする” といった事象を物体と紐付けて理解することができる．

マルチモーダル参照解析の中でも，メンションから物体への直接参照を特定するタスクはフレーズグラウンディングとよばれ [6, 7]，画像とそのキャプションを解析対象としたモデルが提案されてきた [8, 9, 10]．しかし，これらのモデルは間接参照を含むマルチモーダル参照解析を原理的に扱えず，日本語対話テキストのフレーズグラウンディング性能も十分でない [2] [11]．その原因は実世界での対話や

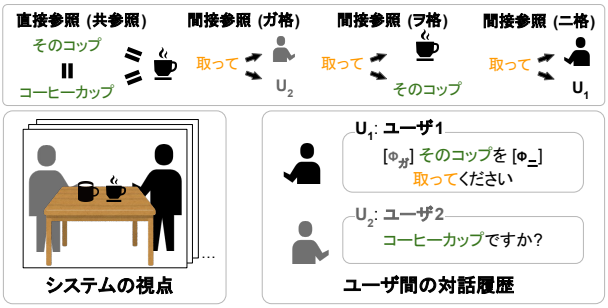

図 1　2 者の実世界対話をシステムが理解する状況におけるマルチモーダル参照解析の例．述語 “取って” におけるガ格とニ格の項は省略されている．

日本語特有の曖昧性にあり，“それ” といった指示詞の発生は前者に，主語や目的語の省略は後者に該当する．我々はマルチモーダル参照解析の性能向上を通してこれら曖昧性の解消を目指す．

本研究では，日本語対話に対するマルチモーダル参照解析の性能向上に寄与する要素について示唆を与える．共参照・述語項構造解析タスクに代表される日本語のテキスト解析の研究では，複数の照応関係[3]を統合的に解析することが，個別のタスク性能の向上に寄与することが示されてきた [12, 13, 14]．我々はこの利得がマルチモーダル参照解析でも得られるかを検証するため，メンション間の照応関係，およびメンション・物体間の参照関係を統合的に解析するフレームワークを提案する[4]．以後，メンション間の解析を照応解析と呼び，メンション・物体間の解析を参照解析と呼ぶ．本研究では提案フレームワークを用いて以下の内容を議論する．

- 照応解析が参照解析に与える影響
- 直接参照・間接参照の同時学習の効果

＊　現所属：日本電気株式会社

1　共参照関係のほか，述語からその項への格関係 [5] などの間接的な関係を含む．メンションから物体への参照解析では．前者・後者の参照を直接参照・間接参照と呼ぶ．

2　例えば，GLIP [9] を用いて日本語対話のフレーズグラウンディングを解く場合，日本語キャプションを解く場合と比較して，Recall@1 に 0.385 ポイントの差が生じる (表 1 参照)．

3　メンション間の参照関係を意味する．

4　https://github.com/SInadumi/mmrr で公開している．

2 準備

本研究は日本語の 2 者対話におけるマルチモーダル参照解析タスク，およびそのデータセットである J-CRe3[5] [1] を対象とする．

2.1 マルチモーダル参照解析

対話テキスト $\mathbf{T} = \{m_1 \cdots m_a \cdots m_b \cdots\}$ とその発話区間に対応する動画のフレーム系列 $\mathbf{V} = \{\mathbf{I}_1 \cdots \mathbf{I}_c \cdots\}$ が与えられた時，J-CRe3 におけるマルチモーダル参照解析では，テキストに含まれるメンション m_b とその参照先に存在する参照関係を特定する．本タスクは，メンション間の照応を解析する照応解析とメンション・物体間の参照を解析する参照解析の 2 つから構成される．

本研究では，参照関係の集合を L で示し，共参照・直接参照の関係 (=) と出現頻度の多い 5 種の格・間接参照の関係を解析する[6]．

照応解析　$\mathbf{T}$ が与えられた時，照応解析では，メンション m_b と照応関係にあるメンション m_a を特定する．m_b が照応先を持たない場合も存在する．

参照解析　$\mathbf{T}$ と $\mathbf{V}$ 中の任意の画像 $\mathbf{I}_c$ が与えられた時，参照解析では，メンション m_b と参照関係にある物体を選択する．具体的には，物体検出モデルにより，$\mathbf{I}_c$ に対して最大 q 個の物体矩形 $\mathbf{O}$ と物体特徴の系列 $\mathbf{X}$ のタプル $(\mathbf{O}, \mathbf{X}) = \{(o_1, \mathbf{x}_1) \cdots (o_q, \mathbf{x}_q)\}$ を推定する．$\mathbf{T}$ と $\mathbf{X}$ を入力として，$\mathbf{O}$ の要素を選択する．ある対話テキストの発話の開始・終了時刻の区間から動画を 1 秒単位で切り出したフレーム系列を $\mathbf{V}$ と見做す．$\mathbf{X}$ の抽出は先行研究 [15, 16] に従う．

2.2 J-CRe3

J-CRe3 は主人とお手伝いロボットの協働作業で発生する実世界対話を対象とし，ロボット視点の動画，俯瞰視点の動画，両者の対話音声を収録したデータセットである．注釈として，対話音声の書き起こしおよび照応・参照関係が付与されている．本研究では照応・参照関係と物体矩形の注釈が付与された，ロボット視点の動画と対話テキストを用いて，照応・参照解析を行う．J-CRe3 では，ある述語から物体への間接参照が存在して，その物体に対応するメンションが無い事例が存在する．この事例は述語から物体へのゼロ参照とよばれる [1]．

5 J-CRe3: A **J**apanese **C**onversation Dataset for **R**eal-world **R**eference **R**esolution

6 L = {=, ガ格, ヲ格, ニ格, デ格, ノ格 } である．

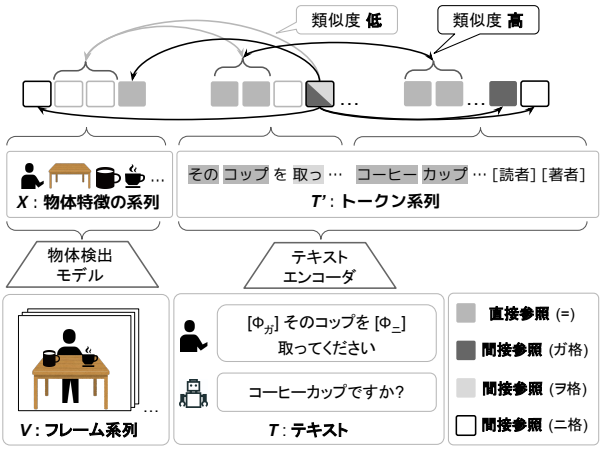

図 2　J-CRe3 の解析における提案フレームワークの概要．照応解析モデルは $\mathbf{T}'$ を用いて照応関係を出力し，参照解析モデルは $\mathbf{T}'$ と $\mathbf{X}$ を用いて参照関係を出力する．

3 提案手法

本研究で提案する，マルチモーダル参照解析のためのフレームワークの概要を図 2 に示す．照応解析モデルと参照解析モデルは独立に学習するが，テキストエンコーダ [17, 18] の重みは共有し，参照関係の統合的解析を実現する．

照応解析モデル　図 2 における我々の照応解析モデルは，メンション同士の埋め込みの類似度から照応関係を解析する．テキスト $\mathbf{T}$ が与えられた時，テキストエンコーダは最大系列長 p，次元 d_T のトークン系列 $\mathbf{T}' \in \mathbb{R}^{p \times d_T}$ を出力する．[19] に倣って，学習可能な重み $\mathbf{W}_{T1} \in \mathbb{R}^{d_T \times d_T \times |L|}$ により $\mathbf{T}'$ に次元を追加し，$\widehat{\mathbf{T}}$ を得る (式 1)．$\widehat{\mathbf{T}}$ 同士の内積を照応関係ごとに算出し (式 2)，類似度行列 $\mathbf{S}_l$ を用いてメンション m_b と照応関係にあるメンション m_a を選択する．埋め込み同士の類似度を用いることで，参照解析と同一の枠組みで照応解析を扱うことができる．

$$\widehat{\mathbf{T}} = \mathbf{T}'\mathbf{W}_{T1} \in \mathbb{R}^{p \times d_T \times |L|} \tag{1}$$

$$\mathbf{S}_l = \widehat{\mathbf{T}}_l \widehat{\mathbf{T}}_l^{\top} \in \mathbb{R}^{p \times p}, l \in L \tag{2}$$

メンションは基本句単位[7]であるが，$\mathbf{T}'$ はサブワード単位である．そこで，メンションの先頭のサブワードをその代表とみなして学習・推論を実施する．この一連の処理は後述する参照解析モデルにおいても同様である．

参照解析モデル　図 2 における我々の参照解析モデルは，メンションの埋め込みと物体特徴の系列 $\mathbf{X}$ の類似度から参照関係を解析する．テキスト $\mathbf{T}$ とフレーム系列 $\mathbf{V}$ の要素である画像 $\mathbf{I}_c$ が与えられた時，

7 1 つの自立語とその前後に付く付属語から成る単位．

表 1　**左**: J-CRe3 における直接参照の解析精度．開発・評価セットを用いた．**右**: Flickr30k-Ent-JP の評価セットにおける直接参照の解析精度．† が付与されたモデルでは，参照解析の学習・評価をランダムシードで 3 回繰り返した平均値を報告する．丸括弧内の数字は正例の数を意味する．

モデル	全体 (996 件)			普通名詞 (671 件)			指示詞 (120 件)		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ベースライン †	0.336	0.567	0.658	0.350	0.573	0.671	0.302	0.575	0.680
w/ CAModel †	0.335	0.579	0.678	0.329	0.558	0.664	0.325	0.636	0.711
w/ KWJA †	0.301	0.536	0.646	0.311	0.551	0.671	0.300	0.544	0.647
w/ Ours †	0.354	0.592	0.688	0.354	0.581	0.674	**0.366**	0.675	0.772
GLIP	**0.437**	**0.695**	**0.748**	**0.433**	**0.679**	**0.733**	0.316	**0.716**	**0.775**

モデル	R@1	R@5	R@10
ベースライン †	0.558	0.735	0.767
w/ CAModel †	0.560	0.734	0.767
w/ KWJA †	0.559	0.733	0.767
w/ Ours †	0.560	0.733	0.767
GLIP	**0.822**	**0.951**	**0.970**

テキストエンコーダと物体検出モデル [20] を用いて，$\mathbf{T}'$ と $(\mathbf{O}, \mathbf{X})$ を構成する．単一の線形層により，$\mathbf{T}'$ の次元 d_T と $\mathbf{X}$ の次元 d_O を d_S に揃える．式 1 と同様，学習可能な重み $(\mathbf{W}_{T2}, \mathbf{W}_O) \in \mathbb{R}^{d_S \times d_S \times |L|}$ を用いて，$\mathbf{T}'$ から $\widehat{\mathbf{T}}$ を，$\mathbf{X}$ から $\widehat{\mathbf{X}}$ を得る (式 3, 4)．ここで，$Dec(\cdot)$ は 2 層の Decoder ブロックであり，$\mathbf{T}'$ による $\mathbf{X}'$ の条件付けを cross-attention で実現する研究 [10, 21, 22] に倣っている．$\widehat{\mathbf{T}}$ と $\widehat{\mathbf{X}}$ の内積を参照関係ごとに算出し，類似度行列 $\mathbf{U}_l$ を用いてメンションから $\mathbf{O}$ の要素を選択する (式 5)．

$$\widehat{\mathbf{T}} = \mathbf{T}'\mathbf{W}_{T2} \in \mathbb{R}^{p \times d_S \times |L|} \tag{3}$$

$$\widehat{\mathbf{X}} = Dec(\mathbf{X}, \mathbf{T}')\mathbf{W}_O \in \mathbb{R}^{q \times d_S \times |L|} \tag{4}$$

$$Dec(\mathbf{X}, \mathbf{T}') \in \mathbb{R}^{q \times d_S}$$

$$\mathbf{U}_l = \widehat{\mathbf{T}}_l \widehat{\mathbf{X}}_l^\top \in \mathbb{R}^{p \times q}, l \in L \tag{5}$$

J-CRe3 における間接参照の注釈の量は限られており，MDETR [8] や GLIP [9] のような学習に大規模な画像・テキストのペアを要するモデルは，J-CRe3 で学習が困難である．そこで本研究では，物体検出モデルの重みは固定し，テキストエンコーダ，およびトークン系列 $\mathbf{T}'$・物体特徴の系列 $\mathbf{X}$ を混合するモジュールのみを学習する．

目的関数　照応解析およびフレーズグラウンディングの先行研究 [9, 14] で使用される，softmax cross entropy を用いる (式 6, 7)．$\mathbf{S}_{(l,ground)} \in \{0,1\}^{p \times p}$ と $\mathbf{U}_{(l,ground)} \in \{0,1\}^{p \times q}$ は，ある照応・参照関係 l における $\mathbf{S}_l$ と $\mathbf{U}_l$ の正例の行列である．

$$\mathcal{L}_S = \sum_{l \in L} loss\{\mathbf{S}_l; \mathbf{S}_{(l,ground)}\} \tag{6}$$

$$\mathcal{L}_U = \sum_{l \in L} loss\{\mathbf{U}_l; \mathbf{U}_{(l,ground)}\} \tag{7}$$

4　実験: 参照解析

本節では，直接参照のみで学習した参照解析モデルの評価結果 (4.1 節)，および間接参照も加えて学習したモデルの評価結果 (4.2 節) について述べる．以後，前者をフレーズグラウンディングの結果と呼び，後者を参照解析の結果と呼ぶ．

実験設定

比較モデル　提案フレームワークの参照解析モデルをベースラインとする．フレーズグラウンディングの比較対象は，その代表的なモデルである GLIP [9] を，日本語データセットで学習したモデルを用いる．参照解析の比較対象は，GLIP と照応解析モデルの KWJA [19] の結果を統合し，これを参照解析の結果としたモデル (GLIP + KWJA) を用いる[8]．

実験では，まず，照応関係の注釈で照応解析モデルを学習する．このテキストエンコーダを参照解析のベースラインの学習で用いることで，共参照解析・照応解析がフレーズグラウンディング・参照解析に与える利得を検証する．照応解析モデルは，3 節のモデル，CAModel [14]，KWJA を用いた．

データセット　フレーズグラウンディングでは，J-CRe3 の直接参照の注釈と Flickr30k-Ent-JP [23] を用いる．参照解析ではこれらに加え，J-CRe3 の間接参照の注釈を追加する[9]．[11] に倣って，GLIP の学習では Visual Genome [6] と GQA [25] で事前学習済みの重みを学習の初期値として使用する．

評価指標　フレーズグラウンディングの評価で使用される Recall@k (R@k, $k = \{1, 5, 10\}$) を用いる．

4.1　フレーズグラウンディングの結果

共参照解析による学習の効果　表 1 に定量評価の結果を示す．J-CRe3 の評価では，共参照解析の学習過程を経た我々の参照解析モデル (ベースライン w/

8　[1, 11] で提案されたモデルである．本モデルではゼロ参照を扱えない．

9　Flickr30k-Ent-JP のキャプションに擬似教師として間接参照の注釈を機械的に付与し [24]，これを学習に用いたが，予備実験の結果，解析精度向上に寄与しないことが判明した．

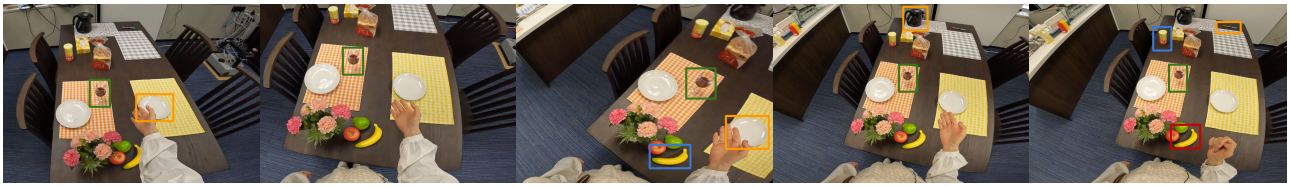

図 3　フレーズグラウンディングの解析の実例．J-CRe3 の評価セットから抜粋 (シナリオ ID: “20230301-57113951-0”)．テキスト中に出現する “これ” と “それ” に対応する正解の物体を緑の矩形で示す．ベースライン，ベースライン w/ Ours, GLIP の **Recall@1 の解析誤り**を，それぞれ青，赤，橙の矩形で示す．

表 2　参照解析モデルの解析精度．J-CRe3 の開発・評価セットを用いた．† と丸括弧は表 1 のキャプション参照．

モデル	直接参照 (996 件)			ガ格 (2,053 件)			ヲ格 (915 件)			ニ格 (1,074 件)			デ格 (139 件)			ノ格 (163 件)		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ベースライン †	0.325	0.575	0.672	0.547	0.732	0.761	0.241	0.517	0.618	0.539	0.734	0.756	0.134	0.280	0.366	0.359	0.556	0.635
w/ CAModel †	0.307	0.537	0.665	0.552	**0.763**	**0.809**	0.240	**0.520**	**0.628**	**0.562**	**0.774**	0.810	0.179	0.287	0.438	**0.447**	**0.662**	**0.717**
w/ KWJA †	0.316	0.569	0.664	0.548	0.745	0.782	**0.249**	0.508	0.604	0.558	0.754	0.784	0.119	0.258	0.366	0.402	0.603	0.664
w/ Ours †	0.322	0.556	0.657	0.548	0.758	0.801	0.236	0.508	0.613	0.555	0.772	**0.812**	0.131	0.318	0.438	0.368	0.576	0.660
GLIP + KWJA	**0.434**	**0.693**	**0.744**	0.063	0.109	0.116	0.177	0.346	0.375	0.063	0.082	0.086	**0.316**	**0.510**	**0.539**	0.239	0.276	0.288

Ours) がベースラインの性能を上回り，統合的解析に利得が存在することを明らかにした．特に，共参照解析は指示詞の解析精度に寄与し，その精度は GLIP に比肩する．一方，Flickr30k-Ent-JP の評価では共参照解析の利得は存在しない．GLIP と異なり，我々のモデルは物体検出モデルの重みを固定しているため，物体検出の精度の上界が定まっていることに起因すると推察する．

定性評価　図 3 に指示詞の解析の実例を示す．我々のモデル (ベースライン w/ Ours) は，ベースライン・GLIP と比較して，指示詞 (“これ”，“それ”) に対する Recall@1 の解析誤りが少ない．一方，図 3 の最終フレームにおいて，我々のモデルは “それ” に対し “バナナ” を推定していることから，フレーム間の予測の整合性に課題が存在することが判明した．

4.2　参照解析の結果

直接参照と間接参照の同時学習の効果　表 2 に定量評価の結果を示す．表 1 と表 2 におけるベースラインの結果より，J-CRe3 の間接参照の注釈を学習に加えると，直接参照の精度 (Recall@5,10) が向上し，直接・間接参照の同時学習はフレーズグラウンディングで有効であることが示唆された．一方，照応解析の学習過程を経た我々のモデルは，直接参照の解析精度が低下している．この結果は，照応解析により，フレーズグラウンディングを解くためのテキストエンコーダの表現力が低下したことに起因すると推察する．照応解析の結果との関連性は C 節を参照されたい．

照応解析による学習の効果　表 2 より，照応解析の学習過程を経たモデルは，照応解析モデルに依らず，間接参照の精度 (Recall@5,10) がベースライン・GLIP + KWJA を上回っており，照応解析による利得が示唆された．例外はヲ格とデ格の結果であり，前者は利得が確認されず，後者は GLIP + KWJA の性能が最も高い．GLIP + KWJA はゼロ参照を扱えないことを鑑みれば，J-CRe3 におけるデ格はゼロ参照の頻度が少ないと言える．

5　おわりに

本研究では，実世界における日本語対話の曖昧性解消を目的として，参照関係を統合的に解析するフレームワークを提案し，参照解析の性能向上に寄与する要素について示唆を与えた．

実験の結果，共参照解析・照応解析の学習過程はフレーズグラウンディング・参照解析の性能向上に寄与することが判明し，参照解析における直接参照と間接参照の同時学習の有効性が示唆された．とりわけ，共参照解析を経た我々のモデルは，指示詞といった曖昧なメンションに対する精度の向上が顕著であった．これらの知見は，実世界で動作する対話システムを実現する際の一つの指針となりうる．

照応関係の学習が参照解析に有効な理由とヲ格で利得が得られなかった理由については，調査を進め，本研究で得た知見の一般化可能性は別モデルやデータで検証する．また，解析精度自体も改良の余地があるため，直接参照・間接参照の注釈に対するデータ増強やアーキテクチャの工夫が必要である．

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 JST さきがけ (JPMJPR24TC) の助成を受けて実施した．

参考文献

[1] Nobuhiro Ueda, Hideko Habe, Akishige Yuguchi, Seiya Kawano, Yasutomo Kawanishi, Sadao Kurohashi, and Koichiro Yoshino. J-CRe3: A Japanese conversation dataset for real-world reference resolution. In **LREC-COLING**, pp. 9489–9502, 2024.

[2] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In **EMNLP**, pp. 4903–4912, 2021.

[3] Te-Lin Wu, Satwik Kottur, Andrea Madotto, Mahmoud Azab, Pedro Rodriguez, Babak Damavandi, Nanyun Peng, and Seungwhan Moon. SIMMC-VR: A task-oriented multimodal dialog dataset with situated and immersive VR streams. In **ACL**, Vol. 1, pp. 6273–6291, 2023.

[4] Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. What you see is what you get: Visual pronoun coreference resolution in dialogues. In **EMNLP-IJCNLP**, pp. 5123–5132, 2019.

[5] Charles J Fillmore. The case for case. **Universals in Linguistic Theory**, pp. 21–119, 1968.

[6] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. **IJCV**, Vol. 123, No. 1, p. 32–73, 2017.

[7] 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.

[8] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR - modulated detection for end-to-end multi-modal understanding. In **ICCV**, pp. 1780–1790, 2021.

[9] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In **CVPR**, pp. 10965–10975, 2022.

[10] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying dino with grounded pre-training for open-set object detection. In **ECCV**, pp. 38–55, 2025.

[11] 植田暢大, 波部英子, 松井陽子, 湯口彰重, 河野誠也, 川西康友, 黒橋禎夫, 吉野幸一郎. J-CRe3: 実世界における参照関係解決のための日本語対話データセット. 自然言語処理, Vol. 31, No. 3, pp. 1107–1139, 2024.

[12] Tomohide Shibata and Sadao Kurohashi. Entity-centric joint modeling of Japanese coreference resolution and predicate argument structure analysis. In **ACL**, Vol. 1, pp. 579–589, 2018.

[13] Hikaru Omori and Mamoru Komachi. Multi-task learning for Japanese predicate argument structure analysis. In **NAACL**, Vol. 1, pp. 3404–3414, 2019.

[14] Nobuhiro Ueda, Daisuke Kawahara, and Sadao Kurohashi. BERT-based cohesion analysis of Japanese texts. In **COLING**, pp. 1323–1333, 2020.

[15] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In **CVPR**, pp. 2117–2125, 2017.

[16] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In **CVPR**, pp. 6077–6086, 2018.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **NAACL-HLT**, Vol. 1, pp. 4171–4186, 2019.

[18] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In **ICLR**, 2021.

[19] Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. KWJA: A unified Japanese analyzer based on foundation models. In **ACL**, Vol. 3, pp. 538–548, 2023.

[20] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In **ECCV**, pp. 350–368, 2022.

[21] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In **ECCV**, pp. 752–768, 2020.

[22] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Semi-supervised multimodal coreference resolution in image narrations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **EMNLP**, pp. 11067–11081, 2023.

[23] Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. A visually-grounded parallel corpus with phrase-to-region linking. In **LREC**, pp. 4204–4210, May 2020.

[24] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. **Computational Linguistics**, Vol. 20, No. 4, pp. 507–534, 1994.

[25] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In **CVPR**, pp. 6700–6709, 2019.

[26] Rahul Aralikatte, Mostafa Abdou, Heather C Lent, Daniel Hershcovich, and Anders Søgaard. Joint semantic analysis with document-level cross-task coherence rewards. In **AAAI**, Vol. 35, pp. 12516–12525, 2021.

[27] Juntao Yu and Massimo Poesio. Multitask learning-based neural bridging reference resolution. In **COLING**, pp. 3534–3546, 2020.

[28] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. arXiv:2310.11441, 2023.

[29] OpenAI. GPT-4o system card. arXiv:2410.21276, 2024.

[30] 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. 213–247, 2014.

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In **ICCV**, pp. 10012–10022, 2021.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In **NIPS**, Vol. 28, pp. 91–99, 2015.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **ICLR**, 2019.

[34] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. BoT-SORT: Robust associations multi-pedestrian tracking. arXiv:2206.14651, 2022.

A 関連研究

照応解析における同時学習の利得 日本語における照応解析の研究では，共参照関係と述語項構造 (ガ格，ヲ格，ニ格)，および用言・体言の述語項構造の同時学習が述語項構造解析に有効であることが，それぞれ，双方向 RNN ベースのモデル [12] や BERT ベースのモデル [13] により示されてきた．英語においても，共参照解析と意味役割付与または橋渡し照応 (ノ格) 解析，の同時学習が双方の精度向上に寄与すると言われる [26, 27]．

一方で，BERT ベースのモデルを提案した日本語の先行研究 [14] では，“共参照関係と格関係の同時学習は互いの精度を低下させる” 報告がなされている． 4.2 節では，“直接参照と間接参照の同時学習は直接参照に有効である” 示唆を得た．これらを鑑みれば，照応解析における同時学習の知見は参照解析に一般化できないと予想できる．

フレーズグラウンディングモデル モデルアーキテクチャは，物体検出モデルの重みを固定して，検出結果の矩形を擬似教師とするタイプ [21, 22] と，物体検出をモデルの学習に組み込み，入力テキストに合わせて動的に矩形を検出するタイプ [8, 9, 10] の 2 種に大別できる．前者は弱教師ありフレーズグラウンディングモデルと呼ばれ，これらは学習コストとモデル精度にトレードオフの関係がある．

提案フレームワークの位置付け 参照解析モデルは弱教師ありフレーズグラウンディングモデルをベースとしており，照応解析モデルはその構成に合わせ，統合的解析を実現した．実験で用いた GLIP+KWJA のように，フレーズグラウンディングモデルと照応解析モデルを組み合わせる手法 [1, 11] と異なり，我々の参照解析モデルはゼロ参照を考慮できる．また，画像に対する prompting 手法 [28] を用いれば，大規模言語モデルに基づく視覚・言語モデル (VLM) [29] でも，我々の参照解析モデルと同等の解析が実現できる．VLM と提案フレームワークの比較は今後の課題としたい．

B 実装詳細

照応解析 テキスト $\mathbf{T}$ の先頭と末尾に特殊トークン ([CLS]・[SEP]) を加える．外界照応の特殊トークン ([著者]・[読者]・[不特定: 人])，メンション m_b に対するメンション m_a が存在しないことを示す特殊トークン ([NULL]・[NA]) も末尾に加える．ここで，[著者]・[読者] は J-CRe3 における主人・ロボットに対応しており，[NULL]・[NA] はそれぞれ，照応解析における格関係の解析と共参照解析で用いる．

テキストエンコーダとして日本語 DeBERTa-v2-large[10] を用いた．トークン系列 $\mathbf{T}'$ の最大系列長は $p = 256$ であり，埋め込みの次元は $d_T = 1,024$ である．特殊トークンを含む $\mathbf{T}'$ の長さが p に満たない場合は zero-padding を行う．

学習には，J-CRe3 に加えウェブ文書・Wikipedia・ユーザ投稿に照応関係の注釈を [30] の手法で付与したデータを混合し，このコーパスを用いた．1 学習事例の単位は連続する 3 文のスパンを 1 文ずつ移動させた事例であり，評価指標は F 値である．

参照解析 物体検出モデルは Detic [20] を用い，Swin-Transformer [31] を backbone としたモデルの重み[11] を使用した．物体矩形 O の予測候補の最大値について，J-CRe3 は $q = 128$，Flickr30k-Ent-JP は $q = 256$ とした．物体検出モデルのモジュールである Region Proposal Network [32] から得られる特徴量を pooling し [15, 16]，最大系列長 q の物体特徴の系列 $\mathbf{X}$ を取得する．実験では $d_T = d_S = d_O = 1,024$，$p = 256$ とした．また，J-CRe3 は 3 文のスパンを 1 文ずつ移動させた事例と 1 画像のペアを 1 学習事例とし，Flickr30k-Ent-JP は最大 5 つのキャプションと 1 画像のペアを 1 つの学習・評価の事例とした．

正例の行列 $\mathbf{U}_{(l,ground)}$ の構成について，J-CRe3 は物体クラスの注釈が付与されているが，Flickr30k-Ent-JP は付与されていない．そこで，あるメンションに対する正解の矩形と予測した矩形 O の要素間で Intersection-over-Union (IoU) を計算し，IoU が 0.5 以上の要素を正例と判断した．

実験設定 照応解析モデルと参照解析モデルの最適化アルゴリズムは AdamW [33] を用い，学習率は 5e-5，weight decay は 0.01，warmup steps は 1,000 とした．epoch 数と batch size はともに 16 である．GLIP のテキストエンコーダは多言語 DeBERTa-v3-base[12] を用い，ハイパーパラメータは [11] に従う．ただし，batch size のみ 16 とした．

J-CRe3 の評価方法 対話履歴の情報を推論で考慮するため，J-CRe3 にから 3 発話のスパンを 1 発話ずつ移動させた事例を 1 評価事例とし，スパンの最終発話 utt_t とそれに対応する動画の 1 フレームを全モデルの評価で考慮した．utt_t を評価する際の動画 $\mathbf{V}$ は，utt_{t-1} の終了時刻から utt_t の終了時刻までの区間を用いた．また，$\mathbf{V}$ のモーションブラーの影響を抑えるため，モデルから得られた物体矩形 $\mathbf{O}$ は，事後的に Detic と複数物体追跡モデル[13] [34] を用いて補正した．

表 3 照応解析モデルの精度．J-CRe3 の評価セットにおける，共参照 (Coref.)，述語項構造 (PAS)，橋渡し照応 (Bridging) の解析結果を報告．◊ は共参照解析を単体で学習した結果 (4.1 節) を意味する．

モデル	Coref. ◊	Coref.	PAS	Bridging
CAModel	0.647	0.613	0.826	0.644
KWJA	**0.666**	**0.637**	**0.837**	**0.720**
Ours	0.616	0.629	0.822	0.688

C 実験: 照応解析

モデル間の精度比較 表 3 に， 4.1 節と 4.2 節で用いた照応解析モデルの解析精度を示す．共参照解析を単体で学習した場合と照応解析を学習した場合，いずれにおいても KWJA の精度が最も良く，我々の提案した照応解析モデルと CAModel は同性能であった．KWJA や CAModel と異なり，提案モデルは参照解析モデルとの共有部分を最大化するため，FFN を用いていない．この違いにより，提案モデルの精度低下が生じたと考えられる．

参照解析との関連性 表 1，表 2，表 3 より，我々の照応解析モデルは照応解析の精度が最も高くないが，これをフレーズグラウンディング・参照解析で用いると双方の精度が向上する．照応解析が参照解析に与えた恩恵の解明も今後の課題であり，BERT の埋め込み表現の変化や類似度行列 $\mathbf{S}$ と $\mathbf{U}$ の観察が必要である．

10 ku-nlp/deberta-v2-large-japanese
11 Detic_C2_SwinB_896_4x_IN-21K+COCO
12 microsoft/mdeberta-v3-base
13 kaiyangzhou.github.io/deep-person-reid/MODEL_ZOO の osnet_x0_75 を用いた．