ホワイトボードを用いた数学協調作業の対話構造解析に向けて

胡真瑜 [1]　朝倉卓人 [2]　吉野幸一郎 [1]
[1] 東京科学大学　[2] 国立情報学研究所
hu.c.873d@m.isct.ac.jp, takuto@nii.ac.jp,
koichiro@c.titech.ac.jp

概要

数学を用いる協調作業対話では、参加者はホワイトボードのような視覚的作業空間を共有しながら、音声言語によってコミュニケーションを行う。このような状況におけるマルチモーダル相互作用を理解する上では、参照関係を含む対話構造解析が重要である。これは例えば、不十分に特定された表現（例えば「これ」や「それ」といった代名詞）とボード上に書かれた対応する数学記号や方程式との間の参照関係などを含む。この課題に対処するため、また Visually Grounded Dialogue の研究から着想を得て、本研究では数学協調作業におけるマルチモーダル参照解析に焦点を当てる。数学表現が説明の過程で段階的に書き加えられ、かつ発話から同時に参照される状況として、数学教育動画を用いる。発話をホワイトボード上の視覚的数学要素へ明示的に接続する画像ベースのアノテーションスキーマを提案し、複数の粒度レベルで参照構造を捉える。本研究ではこうしたホワイトボードを用いた対話構造解析のための基盤を提供し、数学協調作業対話における曖昧性と参照解決の将来的なモデリングを支援することを目的とする。

1　はじめに

数学の問題解決対話では、参加者はホワイトボード上の手書き記号や方程式といった視覚的作業空間を共有し、音声言語によってコミュニケーションを行う。このようなマルチモーダル相互作用を理解する上で重要な課題は、対話発話、特に「これ」や「それ」といった曖昧な代名詞と、それに対応する視覚的な数学要素との間の参照関係を解決することにある。これらの関係を含む対話構造の解析を行うことは、人間の問題解決行動を理解するうえで不可欠であり、数学コンテンツと相互作用する知的マルチモーダルシステムの開発にとっても重要である。

Visually Grounded Dialogue に関する先行研究 [1, 2] では、言語と自然画像の対応付けが検討されてきた。ただしこれをホワイトボード上の数学協調作業の文脈に適用した場合、いくつかの固有の課題が浮かび上がる。数学における参照は、抽象的な記号表現を対象とすることが多く、その意味は視覚的外観だけでなく、数学的構造や談話文脈にも依存する。既存の枠組みは、数学協調作業におけるこのような動的で多段階の参照を扱うことが難しい。

本研究では、こうした数学協調作業の問題解決における曖昧性と参照解決、またの将来的な対話構造解析を支えるため、タスク定義とアノテーションガイドラインを作成する。本研究ではアノテーション対象となる題材として、数学教育動画を利用する。こうした動画は講義形式であり、対面の相互対話より構造化されているが、代名詞の使用や視覚に基づく言及といった数学協調作業における中核的な参照現象を含む。本研究では、発話を視覚的数学要素に接続する画像ベースのアノテーションスキーマを設計し、粒度の異なる複数レベル（数式、スパン、項）で参照関係を捉える。本研究は、数学の問題解決対話におけるマルチモーダル参照解決の枠組みを提案し、対話構造におけるギャップを埋めることを目指す。

2　関連研究

参照解決に関する研究は、主としてコンピュータビジョンと数学情報検索のコミュニティで、それぞれ別の流れとして発展してきた。数学協調作業対話におけるマルチモーダル参照解析を扱う本研究は、これら領域の交差点に位置付けられる。

対話における視覚的グラウンディング　Vision and language の研究では、参照解析は通常、自然言語表現を視覚シーン内の領域へ接続する問題として定式化される。近年多くの先行研究が存在し [3] 、特に静止画像中の具体的物体への参照グラウンディ

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

ングで高い性能が示されてきた。より最近の研究では、言語的文脈とマルチモーダル文脈を同時にモデル化することで対話や現実世界の相互作用へ拡張したり [1]、これをベンチマークとする J-CRe3 [2] のようなデータセットが提案されたりしている。しかし、これらは主として比較的状態変化が少ない日常物体を対象としている。これに対して数学協調作業対話ではホワイトボードの記述に代表される視覚的作業空間が動的に変化し、問題解決中に表現が段階的に書かれ、参照される。

数学記号処理　数学記号処理、特に数学情報検索の研究では、テキストベースの記号処理に焦点が当てられてきた。Symlink [4] のような共有タスクは文書内の記号共参照を扱う。いくつかの研究では、数学記号は単一テキスト内であっても頻繁な再定義や文脈依存性のために非常に曖昧であることが示されている [5]。こうした曖昧性解消には談話レベルの文脈が重要である点が議論されており [6]、数学的関心対象（Mathematical Objects of Interest; MOIs）[7] の概念とも密接に関連している。手書き数式認識（例えば image-to-LaTeX 変換 [8]）は広く研究されているが、これらの手法は通常、表現を独立した静的入力として扱い、対話における参照関係はモデル化しない。

視覚と数学の境界領域研究　視覚的グラウンディングと数学記号処理の交差領域における参照解決は十分に研究されていない。特に、不十分に特定された対話表現が、動的に変化する視覚的数学コンテンツにどのようにグラウンディングされるかを検討した先行研究は稀である。本研究ではこのギャップに対処し、複数の粒度レベルで対話表現を視覚的数学要素へリンクするアノテーション枠組みを提案する。

3　タスク定義

本研究では、数学協調作業におけるマルチモーダル参照解析タスクを定義する。入力はホワイトボードセッションを記録した動画と参加者の発話に時間同期された書き起こしであり、出力は対話中の言語表現とホワイトボード上に書かれた対応する数学的対象との参照関係のラベルである。出力はグラウンディングされた対 (m, b) の集合であり、m は書き起こし中のテキストスパン（例えば「これ」のような代名詞、あるいは名詞句）を表し、b は発話時点の動画フレーム内で参照される数学記号、部分表現、または方程式を同定するバウンディングボックスである。

図 1 にこのタスクの例を示す。対話の中で講師が「this bottom part is secant squared」と述べた場合、テキストスパン「this bottom part」をホワイトボード上の手書き表現 sec^2u を含むバウンディングボックスに結び付ける。また、その後で講師が「two of them」と述べたとき、以前に言及された式の一部である対応する項（term）を特定する必要がある。一般的な共参照解析とは異なり、ここでの参照対象は、数秒あるいは数分前に書かれた可能性のある視覚的オブジェクトであり、音声言語と視覚コンテンツのクロスモーダルな対応付けが要求される。

静的な視覚シーンにおける参照解決（例：VQA）と比べ、このタスクにはいくつかの特徴がある。第一に、視覚文脈は動的であり、数学表現は段階的に書かれたり修正されたりし、動画の時点によっては参照候補が存在しないことがある。第二に、参照の粒度が多様であり、個々の項（term）から部分表現のスパン（span）、完全な式（equation）まで対象となる。第三に、参照は時間的局所性が強く、一般に現在書いている表現や直前に言及された表現が参照対象となりやすい。

4　データセット

本研究ではアノテーションスキーマを構築するため、公開されている数学教育動画（例：YouTube や大学の OpenCourseWare）を用いる。動画は主として高校レベルおよび学部初年次で扱われる基礎数学（例：微積分、線形代数）を対象とする。約 100 本規模の動画を取得してアノテーションすることを目指しており、これは代表的な参照現象を十分にカバーしつつ、人手アノテーションとして実行可能な規模であると見込まれる。

教育動画を選ぶ理由は、最終的な数学協調作業対話の解析に必要な要素を含み、かつ比較的制御された状況で収録されたデータであることが期待されるためである。これらの動画には、音声による参照解析、視覚に基づく言及、数学表現の段階的構築といった、本タスクに必要な本質的なマルチモーダル参照現象が含まれている。さらに、この種のデータは豊富で入手容易であり、大規模アノテーションに適している。

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

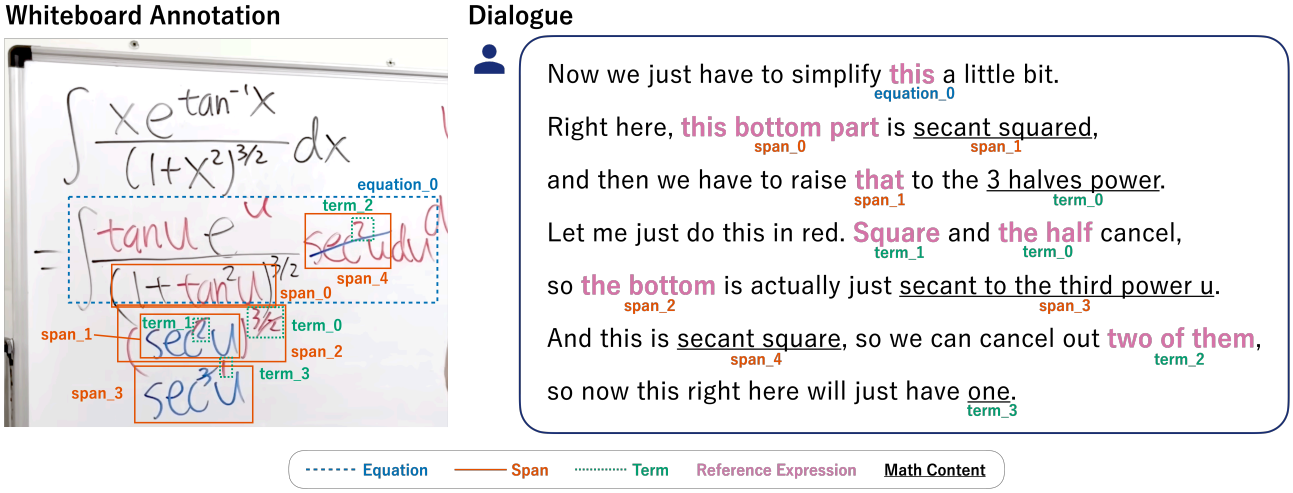

図 1　数学教育対話へのアノテーション例。対話内で明示的に言及された数式のみをアノテーションしている。代名詞、不十分に特定された表現、および発話された数式は、対応する視覚的参照先（グラウンディング先）に紐付けられている。公開動画 [9] の 02:15 時点のスクリーンショット。

5　アノテーション設計

図 1 にアノテーションされた動画の例を示す。談話上の言及が異なる粒度レベルでホワイトボード上の視覚的数学表現にグラウンディングされる様子を示している。本節では、アノテーション範囲、提案スキーマ、そして実際のアノテーション上考慮すべき事項を説明する。

5.1　アノテーション範囲と方針

数学協調作業の特徴に基づき、アノテーションの範囲を限定する。具体的には、ホワイトボード上に新しく書かれ、かつ講師によって明示的に参照された数学表現のみをアノテーションする。参照は、口頭による言及、あるいは明確な視覚的指示によって判断される。可視状態にあっても説明中に言及されない表現は除外する。この方針は、数学協調作業における強い局所性を反映してアノテーション労力を妥当な参照候補に集中し、意味のある参照イベントの被覆率を保ったままコストを大幅に削減することを目的としている。

ジェスチャー情報は現時点ではアノテーションしない。その代わり、画像レベルの視覚単位と談話言及とのリンクに焦点を当て、ジェスチャー手がかりは将来研究で補完的モダリティとして扱うことを検討する。

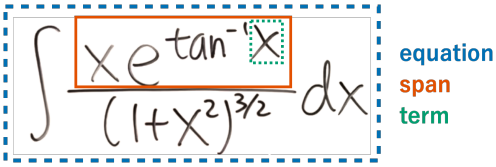

図 2　階層的アノテーションスキーマ。対話と視覚的コンテンツを正確に対応付けるため、**式**（Equation：青の破線）、**スパン**（Span：オレンジの実線）、**項**（Term：緑の点線）という 3 つの粒度レベルを定義している。

5.2　アノテーションスキーマ（式、スパン、項）

複数粒度での参照解析のため、3 種類の視覚単位（式、項、スパン）からなる最小限のアノテーションスキーマを提案する。図 2 に各単位の具体例を、また図 1 にこれを用いたアノテーションの具体例を示す。

式（Equation）：等式、積分、導出された公式など、全体として参照され得る視覚的に完結した数学表現に対応する。**項（Term）**：談話で独立に参照され得る最小の数学単位であり、単独の実体として議論される変数や定数などを含む。**スパン（Span）**：1 つ以上の項または範囲からなる合成で、説明中に単一の指示単位として扱われるものを指す。典型例は $1+x^2$、$tan^{-1}x$、$e^{tan^{-1}x}$ のような部分式である。「この部分」「その式」といった代名詞的表現で参照されることが多い。

スパンは内部に項を含むことがあるが、最小参照単位へのアクセスを確保するため、項は常にアノテーションする。一方で演算子や構造記号（例：積分記号、微分記号）は、教育談話では独立に参照さ

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

れることが稀であるため、単独ではアノテーションしない。このスキーマは数学意味論よりも発話での言及に関連する視覚単位を捉えるよう設計する。

5.3 アノテーション上の考慮事項

上記の方針に従いアノテーションを行った結果、いくつかの課題が浮かび上がった。

第一に、分数、根号、上付き文字のように、参照対象が不規則あるいは非矩形の視覚構造を持つ場合があり、対応範囲を矩形バウンディングボックスで正確に表現するのが難しい。現状は視覚的に顕著な領域を覆う近似としてバウンディングボックスを用いている。将来的には矩形でない領域表現も検討を行う必要がある。

第二に、今回は J-CRe3 用に開発されたアノテーションツールを用いたが、やはり数学協調作業特有の課題が生じた。書き起こし中に代名詞が反復して出現するケースでは、この明示的な曖昧性解消が必要となる。そこで本研究では、軽量なインデックス付け（例：`it_00`, `it_01`）を用いた。また、バウンディングボックスのメタデータに LaTeX 形式で符号化された構造的数式を保存する必要があるため、手動後処理が必要となる場合がある。

最後に、複数の表現が空間的に近接している場合や、段階的筆記の途中で部分式のみが可視である場合、参照同定に曖昧性が生じうる。この曖昧性は、数学協調作業対話の参照解析モデルを構築する際の困難さの一つとなる可能性がある。図 1 に示される span レベル参照や代名詞グラウンディングがその例である。

パイロットアノテーション結果　提案した設計の実現可能性を評価するため、約 6 分の数学教育動画に対して密なパイロットアノテーションを行った。このセッションから、数学的導出の異なるステップに対応する 13 枚のキーフレームを抽出し、そのうち 6 枚を密にアノテーションした。パイロットでは、式、スパン、項を横断して約 40 個のグラウンディング済み視覚エンティティが得られた。また、アノテーションされたキーフレームあたり平均 6〜7 個の参照対象が得られた。このパイロットアノテーションから、特に入れ子構造や段階的筆記における境界曖昧性の扱いを中心に、ガイドラインを改善している。

6 まとめ

本論文では、数学協調作業対話における対話構造解析、特にマルチモーダル参照解析を目的としてタスク定義とアノテーション枠組みを導入した。ホワイトボードなどの共有視覚作業空間上の手書き数学表現に対して、不十分に特定された談話言及をグラウンディングすることに焦点を当てた。アノテーション対象として数学教育動画を用い、式、スパン、項を含む複数粒度の参照構造を捉える最小限の画像ベースアノテーションスキーマを提案した。この枠組みは、数学対話における参照現象を分析するための基盤を提供し、マルチモーダルな問題解決場面における曖昧性と参照解決を支援する。

7 今後の課題

今後は、追加の教育動画および対話参与者同士の問題解決相互作用を取り入れることで、データセットの規模と多様性を拡大し、より幅広い数学トピックと相互作用の分析を可能にすることを目標とする。また、数学的関心対象（MOIs）[5, 7] に関連する中間構造など、複合的な数学単位をより適切に捉えられるよう、アノテーション枠組みを洗練させる。さらに、アノテーション済みデータに基づき、自動参照解決のためのマルチモーダル学習手法を検討する。この際、アノテータ間のばらつきを曖昧性や参照不確実性のシグナルとして活用する可能性も探る。さらに、現在の画像ベース枠組みは、時間ダイナミクスや指差しなどのジェスチャー情報を取り込むことで動画レベル表現へ拡張できる。長期的には、同様のマルチモーダル参照現象が生じる他の技術領域にも一般化することを目標とする。

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は JST さきがけ（JPMJPR24TC）および JST BOOST（JPMJBY24A2）の支援を受けた。

参考文献

[1] Shohei Inadumi, Naoya Ueda, and Koichiro Yoshino. Disambiguating reference in visually grounded dialogues through joint modeling of textual and multimodal semantic structures. In **Proceedings of ACL 2025**, pp. 11183–11198, 2025.

[2] Naoya Ueda, Hiroki Habe, Yusuke Matsui, Akira Yuguchi, Shogo Kawano, Yuki Kawanishi, Sadao Kurohashi, and Koichiro Yoshino. J-cre3: A japanese conversation dataset for real-world reference resolution. In **Proceedings of LREC-COLING 2024**, pp. 9489–9502, 2024.

[3] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. **Journal of Artificial Intelligence Research**, Vol. 71, pp. 1183–1317, 2021.

[4] Viet Lai, Amir Pouran Ben Veyseh, Franck Dernoncourt, and Thien Nguyen. SemEval 2022 task 12: Symlink - linking mathematical symbols to their descriptions. In **Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)**, pp. 1671–1678, Seattle, United States, July 2022. Association for Computational Linguistics.

[5] André Greiner-Petter, Terry Ruas, Moritz Schubotz, Akiko Aizawa, William Grosky, and Bela Gipp. Why machines cannot learn mathematics, yet, 2019. https://arxiv.org/abs/1905.08359.

[6] Takuto Asakura and Yusuke Miyao. What is needed for intra-document disambiguation of math identifiers? In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 17500–17512, Torino, Italia, May 2024. ELRA and ICCL.

[7] André Greiner-Petter, Moritz Schubotz, Felix Müller, Corinna Breitinger, Harvey Cohl, Akiko Aizawa, and Bela Gipp. Discovering mathematical objects of interest—a study of mathematical notations. In **Proceedings of The Web Conference (WWW) 2020**, pp. 1445–1456, 2020.

[8] Yuntian Deng, Anssi Kanervisto, Alexander M Rush, and SEAS Harvard. Image-to-markup generation with coarse-to-fine attention. In **Proceedings of the 34th International Conference on Machine Learning (ICML)**, pp. 980–989, 2017.

[9] bprp calculus basics. Can someone help me do this integral? reddit r/calculus. YouTube video, 2025. https://www.youtube.com/watch?v=mkQp8w17GMs&t=135s.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).