

研究終了報告書

「数式と自然言語の統合的解析による学術文献理解の研究」

研究期間: 2020 年 11 月 ~ 2023 年 3 月

研究者: 朝倉 卓人

加速フェーズ期間: 2023 年 4 月 ~ 2024 年 3 月

1. 研究のねらい

論文や専門書などの科学技術文書から情報抽出を行い、またそのように収集した情報に効率よくアクセスできるようにすることは、自然科学研究の一層の加速・発展に繋がるため重要である。科学技術文書においては、数式が重要な役割を果たすため、自然言語と合わせて数式部分についても詳細な解析を行う必要がある。しかし、プログラミング言語や一階述語論理などの形式的な表現に現れる算式とは異なり、文書中に現れる数式にはさまざまな曖昧性があることがわかっている。本研究では、そうした数式中の曖昧性の解消に取り組むため「数式グラウンディング」と呼ばれるタスクを提案し、その自動化を目指した。

機械学習のアルゴリズムによって得られるのは関数 $y(x)$ である。この関数に、新たに数字の画像 x を入力すると、目標ベクトルと符号化の仕方が等しい出力ベクトル y が出力される。関数 $y(x)$ の詳細な形は訓練データに基づいて求められる。(C.M. ピショップ『パターン認識と機械学習』p. 2)

数学概念

- 関数 $y(\cdot)$
- 出力ベクトル y



図 1: 数式グラウンディング

数式グラウンディングとは簡潔に言えば数式内のトークンをその参照する数学概念(意味)に紐付けるタスクである(図 1)。自然言語中の数式においては 1 種類のトークンが、1 文書内でも複数の意味で用いられる(図の例では y がベクトルと関数の 2 つの意味をもつ)ため、このタスクは一般に容易ではない。また、その達成には数式内のみを解析するのでは不十分で、前後のテキスト(例えば図 3 で波線を引いた「関数」や「出力ベクトル」などの同格名詞)も参考にする必要がある。数式グラウンディングの自動化は、理工系学術文書を計算機に読み込ませ、定理証明支援や数式処理システムなどを用いたさまざまな応用タスクを解く上でなくてはならないステップである。この自動化に取り組むことで、第一に今後の応用に役立つ基盤技術を作ること、第二に自然言語テキスト中の数式という、これまで人類にとってほとんど未知のままに残されていた言語現象について、少しでも多くの知見を掘り起こすことが本研究のねらいである。

加速フェーズでは、数式グラウンディングが解決すべき大きな課題の 1 つである数式トークンの文書内曖昧性を解決する具体的な手法の構築に取り組む。数式トークンの意味の曖昧性自体はこれまでの研究でも認識・調査されてきており、一定の進展が得られている。しかし、文書内に閉じた曖昧性については、曖昧性の解消手法はもちろんその頻度や複雑さ、解消難易度などといった基本的な性質すらも明らかになっていない。これはこれまで数式トークンの文書内曖昧性の調査を行うのに十分な質と量をもつデータ資源が存在していなかったことに起因する。加速フェーズでの研究により、こうした数式トークンの文書内曖昧性が実際に広く存在する複雑で深刻な問題であることを明らかにするとともに、その解消に必要な情報種を特定する。また、このように有効であることが示された情報種を用いて機械学習モデルを訓練し、実際に文書内曖昧性の解消を行うモデルを構築する。これにより、これまで未知のまま残されてきた数式トークンの文書内曖昧性を解消する最初的手法を提案する

とともに、今後の研究においてさらにその解消精度を高めていく上で重要な指針となる知見を得ることを目指す。

2. 研究成果

(1) 概要

本研究では、文書中の数式内に現れる記号(数式トークン)の意味を明らかにする「数式グラウンディング」自動化を目指した。この目標の達成に向けて、十数名の研究協力者とともにこれまでに開発した専用のソフトウェア MioGatto を用いて数式グラウンディング自動化のためのデータセットの構築を行った。その結果、情報学・論理学・数学・天文学などさまざまな学術分野の論文 15 本に対する人手での注釈付け(アノテーション)が完了し、まずは自動化を行うための素地が完成した。またこのデータ構築の過程の中において、人間であれば概ね 90%以上の精度をもって正しく数式トークンの曖昧性を解消し、グラウンディングタスクを解くことができることを明らかにした。

構築したデータセットの分析を通して、数式グラウンディングの自動化に向けた糸口を探った。その結果、自然言語で書かれた文書中の数式トークンの意味が継続する範囲(スコープ)は極めて複雑な様相を呈していることが再確認される一方で、数式トークンの意味が切り替わるポイントを自動的に判別する上で手がかりになる特徴を多数発見した。こうした特徴を最大限活用することで、ルールベースのシンプルなベースライン手法でスコープ切替位置の約 88%を自動的に特定できることを明らかにした。スコープの切替位置をある程度の精度で自動推定できるようになったことで、人手によるアノテーションの作業量を大幅に削減することに成功した。今後はより効率よくデータを拡充することが可能となるため、深層学習などの現代の統計的な手法が適用できる大規模データセットを構築できる環境を創出した。

これまでの成果をまとめた論文は、国際会議 CICM の Math UI ワークショップや、言語資源に関連する研究テーマを専門として扱う言語リソースに関する代表的国際会議である LREC 2022 に採択された。また国内の研究者向けには言語処理学会第 28 回年次大会(NLP2022)にて研究成果の口頭発表を行い、委員特別賞を受賞した。独自に開発した専用のアノテーションツール MioGatto はオープンソースソフトウェアとして一般に公開し、また構築したデータセットも研究者向けに公開を行った。

加速フェーズでは、数式トークンの文書内曖昧性にフォーカスを絞った上で、その具体的な解消手法を提案した。まず、ACT-X 期間中に作成したデータセットを 2 倍以上のサイズに拡張し、論文 40 本から成るデータセットを完成させた。このデータセットを多層パーセプトロンモデルの訓練と評価に用いることで、手法の構築を実施した。具体的には、数式トークンの中でも最も頻出な種類である数式識別子の文書内曖昧性解消を機械学習タスクとして定式化し、その解決に役立つ情報種を特定した。

(2) 詳細

研究項目 (A) 数式グラウンディング自動化

本研究のために開発した専用アノテーションツール MioGatto(図 2)を用いて、のべ 15 名のアノテータとともに複数の学術論文に対するグラウンディング情報(数学概念・グラウンディング情報源)のアノテーションを行った。このデータ構築の過程では、アノテーション作業を肅々

と進行する傍ら、より効率的なアノテーションを行うためさまざまな工夫を凝らし、対応する機能を順次 MioGatto に追加していった。そうした効率化の工夫に関する知見も含め、このツールとデータ構築に関連する新規の学術的貢献は MathUI 2021 ワークショップにて発表を行うとともに、ツール自体もライセンスの下で公開を行った (<https://github.com/wtsnjp/MioGatto>)。

III-A Goals

As illustrated in Fig. 4, in a regression problem, we are given a training set $\mathcal{D}$ of N training points (x_n, t_n) , with $n = 1, \dots, N$, where the variables x_n are the inputs, also known as covariates, domain points, or explanatory variables; while the variables t_n are the outputs, also known as dependent variables, labels, or responses. Note that the outputs are continuous variables. The problem is to predict the output t for a new, that is, as of yet unobserved, input x .

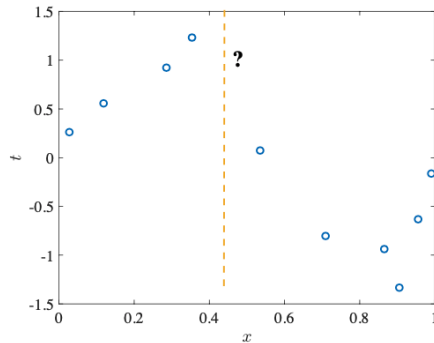


Fig. 4: Illustration of the supervised learning problem of regression: Given input-output training examples (x_n, t_n) , with $n = 1, \dots, N$, how should we predict the output t for an unobserved value of the input x ?

MioGatto v0.4.1 ×

Revision: 6d63b93

Paper ID: 1808.02342

Annotator: Takuto Asakura

Links: [help](#), [bug reports](#)

Options ×

☐ Highlight only relevant SoGs

Progress ×

Concepts: 937/937 (100.00%)

Sources: 232

Concept ×

Please select a math identifier first.

図 2: アノテーションツール MioGatto (スクリーンショット)

上記の取り組みにより、現在までに情報学・論理学・数学・天文学などさまざまな学術分野の論文 15 本に対する完全な人手アノテーションを完了し、数式グラウンディングデータセット (表 1) を得た。特に論文 1 (約 20 ページの長編論文) については 5 名のアノテータによって独立にアノテーションを行い、アノテータ間一致率を計算することで構築したデータの一貫性・再現性を確認した。その結果単純一致率が 84.2%~96.5%、Cohen's κ が 0.75~0.94 の範囲に収まり、このグラウンディングデータセットは人手であればかなり高い品質で構築できることが明らかとなった。このデータセットを分析することで、論文中の数式における記号のスコープが極めて複雑な様相を呈していることや、数式トークンとその情報源の距離は中央値にして 1~2 単語と比較的近い位置に存在するケースが多いことを明らかにした。なお当該のデータセットは、著者が参加する国際的な研究者グループ SIGMathLing のリポジトリにおいて、メンバー限定公開を行った (<https://sigmathling.kwarc.info/resources/grounding-dataset/>)。さらにはドイツの研究グループ KWARC の研究者らとともに W3C Web Data Annotation Model に準拠したグラウンディングデータセットの表現形式を考案した。これにより、当該グラウンディングデータセットは他の研究グループによる別のアノテーションデータと共存し、統一的に取り扱うことが可能となった。

数式グラウンディングデータセット							
論文	分野	単語数	種類	出現	辞書項目	平均候補数	情報源
1	ML	10976	40	937	104	6.4	232
2	NLP	4267	42	266	73	2.6	30
3	NLP	3563	38	433	79	2.5	34
4	論理学	3567	46	1648	64	1.9	30
5	代数学	13154	141	4629	424	5.2	180
6	NLP	2881	25	162	30	2.7	12
7	NLP	5543	31	203	47	2.6	36
8	NLP	4613	23	217	27	1.1	28
9	NLP	6255	34	510	74	2.7	27
10	NLP	5415	73	1175	167	3.3	60
11	NLP	4451	33	237	61	2.9	34
12	NLP	4261	31	186	39	1.7	25
13	NLP	2257	23	124	27	1.2	18
14	天文学	10032	59	1064	129	4.2	97
15	天文学	4863	41	561	73	2.3	95
合計	—	86098	680	12352	1418	—	938

表 1: グラウンディングデータセットの概要

このようにある程度のデータを集めることに成功した後は、自動化に向けた取り組みにも着手した。しかし数式グラウンディングのすべてのプロセスの自動化を行うには依然としてデータが不足しているため、人手では作業に時間のかかる部分を優先的に自動化する方針を採用することとした。具体的には、グラウンディングの第 3 ステップである「数学概念辞書と数式トークン出現の紐付け」を自動で行うべく Major Class Baseline の検討と開発を行った。その結果、既出の数式トークンが現れた際に (1) 原則としては前回出現と同じスコープが継続される、(2) ルールベースで数式トークンの接辞(アクセント記号や添字の有無)を捉えた際に、その内容が前回から変化している場合はスコープ切替が起こった、と判断することで収集した 15 本分の論文データ全体の 88.38% のスコープ継続/切替の区別が行えることが明らかとなった。このベースラインでは捉えることができない残りのスコープ切替位置は、セクションや段落などの文書構造やグラウンディング情報源の有無を基に判定を行う必要があり、そのためのルールや分類手法は、本研究プロジェクトの残り半年の研究期間でさらに分析を進めていく。

研究項目 (B) グラウンディング情報源の分類

研究項目(A)のためのデータ構築を行う中で収集したグラウンディング情報源に対して、新たに定義・宣言・予約のラベル付けを行い、その自動分類に取り組む計画であった。しかし、特に非数学分野の論文においては、本来存在するはずの定義や宣言の区別は明確でなく、それらの一貫したラベル付け自体が困難であることが明らかとなった。

定義・宣言・予約などのラベル付けを行うため、MioGatto を拡張してこのラベル付けを行

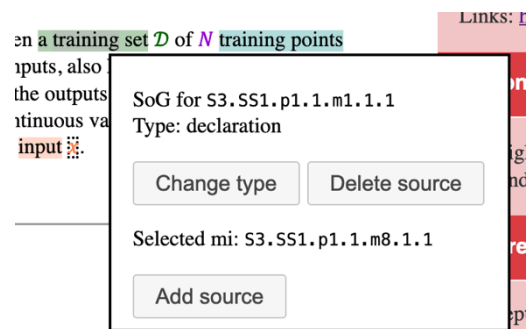


図 3: MioGatto の定義・宣言ラベル付け機能

う機能を開発、完成させた(図3)。しかし、実際に学生アノテータとともに試験的なアノテーションを行ってみると、そもそも本研究で収集した多くのグラウンディング情報源については定義と宣言の具体的な区別を付けることが難しいことが判明した。海外の専門家とも熟議を重ねたが、現時点では一般の科学技術論文に関して、定義・宣言の区別を高いアノテータ間一致率の下で一貫してアノテーションするのは困難であると結論付けざるを得なかった。これらのラベル付けを正確に実施するためには(1)対象論文が数学分野の文書であること、(2)数学を専門とするアノテータがラベル付を行うこと、の2つの条件が必須と考えられる。本研究では数学に限らず多様な理系分野を対象にアノテーションを行ってきたこと、また募集したアノテータも数学を専門とする者はごく少数であったことから、現在の体制では本研究項目についてアノテーションを継続・完了することは困難だと判断した。

研究項目(C) 数式識別子の文書内曖昧性の解消(加速フェーズ)

加速フェーズでは数式識別子の文書内曖昧性を具体的に解消する方法の構築に注力した。まず、機械学習モデルの訓練や評価に十分な量となるようACT-X期間中に作成したデータセットを2倍以上のサイズに拡張した(表2)。その結果、データセットの構成は論文40本となり、その中の数式識別子の出現27,655個に対して文書内曖昧性の解消結果を含む情報がアノテーションされた状態となった。

拡張前後の比較						
	論文	本文の単語	識別子の種類	出現	辞書項目	情報源
拡張前	15	86,098	680	12,352	1,418	938
拡張後	40	237,062	1,742	27,655	3,603	2,466

表2: 拡張前後のグラウンディングデータセット

このデータセットを利用して、多層パーセプトロンモデルの訓練と評価を行った。

3. 今後の展開

本研究を通して、数式グラウンディングの自動化において必須インフラであるアノテーションツール、データセット、そしてベースライン手法を構築・提案した。また構築したデータセットの分析結果や、研究に関する口頭発表に対する世界中の研究者のフィードバックから、数式グラウンディングは確かに自動化が実現可能なタスクであるとともに、理数系学術文書の解析を要する種々の下流タスクを解くにあたって必須となる基盤タスクであることが確認された。今後はこれまでに構築したインフラをもとに、実用レベルの自動化手法を作り上げていくことが課題となる。

実用に耐え得る高精度な自動化手法の獲得にあたっては、これまでに構築したデータセットの分析やテストデータとしての活用を行いながら、個人研究者としてモデル開発を行っていくことも大切であるが、同じ研究コミュニティに属する多様な研究者を巻き込み、精度向上のペースを加速していくことも重要である。具体的には、ベースライン手法や改良手法を用いてさらにリファレンスデータを拡張しつつ、国際的な自動化コンペティションの開催などを通してコミュニティを主導していく活動を行っていく予定である。こうした取り組みにより、もう1-2年のうちにコミュニティとし

て数式グラウンディングの精度向上に取り組む状況を作りたい。

数式グラウンディング手法の精度向上のトレンドが出来上がった後には、実際の応用技術の開発や実装に取り組む必要がある。数式グラウンディングの機能を組版システム LaTeX や論文閲覧ツール(XML やPDF リーダ) 上に実装することにより、そのまま読者支援に役立てることができる。こうした直接的な応用は、精度向上のペースを見極めて 2-3 年のうちにプロトタイプを完成させたい。

さらには後段の応用タスクにつなげる研究にも取り組む必要がある。具体例を挙げると、数式を対象とする検索技術の構築には数式グラウンディングが不可欠である。一般に、数式トークンは 1 文字は数文字程度の短い文字列で表されることが多いため、Google 検索をはじめ現在多くの場面で活用されているキーワード検索は数式に対してはあまり効果的ではない。数式グラウンディングにより検索対象の数式内の各トークンの意味を予め解析しておくことにより、ユーザが見つけたい数式を効果的に見つけることができるような検索システムを構築できると考えられる。arXiv など研究業界に欠かせないサービスに対してはもちろん、民間企業等の内部文書や一般書籍のデータベースにおいても数式検索が必要となる事例は数多く想定されるが、数式グラウンディング手法やその開発過程で得られた知見はこうした場面で役立つと考えられる。実際に、民間企業からそうした案件の相談を受けることもあるため、中長期ではこうしたシステムの開発にプロジェクトリーダー等として関わりたいと考えている。こうした技術の本格的な自動化には、5-10 年単位の時間が必要であると想定している。

(加速フェーズ実施後追記)

加速フェーズの成果により、数式識別子の文書内曖昧性という、数式グラウンディング実現の上で特に重要なステップを一定の精度で具体的に自動化することが可能となった。本研究で構築したモデルは、すでにルールベースの手法を上回り人間アノテータに近い精度を実現しているが、モデル設計はシンプルであり、改善の余地があると考えられる。一方、本研究で見出した曖昧性解消に必要な情報種に関する知見はある程度普遍的な性質であると考えられ、この知見をベースにさらに発展的なアーキテクチャをもつモデルの設計への活用が期待される。今後は、さまざまな機械学習モデルの扱いに長けた研究者も含め、コミュニティ全体を巻き込んでさらなる精度向上を目指していくことが望ましい。加速フェーズ期間中にデータセットを拡張し、数式識別子の文書内曖昧性解消を具体的に機械学習モデルで解決可能な形に定式化したことで、そのための下地は完成している。したがって国際的な自動化コンペティションの開催などの取り組みは、すぐにでも着手が可能であり、当初の見込み通り 1-2 年のうちには進展が得られるだろう。

また、数式グラウンディングの完全な実現には数式識別子の文書内曖昧性以外の問題の解決にも取り組む必要がある。具体的には、文書ごとの数学概念辞書の自動生成や、数学概念と外部知識との紐付けなどが必要となる。これらはそれぞれ数式言語処理分野で研究されてきた定義抽出と、数式・数学的知識の情報抽出(MathIR)に関する研究が利用できると考えられるが、数式グラウンディングの問題設定に合わせた調整や、さらなる精度向上への取り組みを実施する必要がある。これは数式識別子の文書内曖昧性解消のみを解決するよりは時間を必要とすると考えられるが、近年トレンドとなっている大規模言語モデル等も活用することで、2-5 年以内には大きな進展を生むことも可能だろう。

4. 自己評価

本研究課題は、自然言語分野における自動要約や自動翻訳のようにすでに確立されたタスクではなく、数式グラウンディングというまったく新しいタスクを提案・解決しようとするものである。そのため、既存のタスク設計やデータセットを流用することは難しく、タスクの定義からデータセットの構築、評価指標の設計、そしてタスクを解く手法の開発のすべてをゼロから独自に実施する必要があった。そのような高いハードルが存在する中で、アノテータとともに試行錯誤を繰り返しながら、新たなアノテーション手法および高品質のデータセットを完成させ、有名な国際会議で論文を発表するに至った。新規のタスク提案を行う上で、基本となるインフラ整備を短期間で高クオリティに遂行することができたと評価できる。一方、数式グラウンディングの自動化に関しては、実際に構築したデータを詳細に分析することで、ルールベースによるベースラインを構築できたものの、実用レベルに到達するにはもう少し継続した研究努力が必要という結論に至った。データ分析を通じて数式グラウンディングが実際に難しいタスクであることを明らかにしたこと、またどのような特徴量を用いれば、できる限り高い精度でその自動化を達成し得るのか等について知見をもたらすことができたのは前進だが、実応用が可能な自動化手法を完成するには、より長期に継続した研究が必要である。

研究費は、主にデータ作成、研究補助員の雇用、国際的な共同研究の推進に活用した。特に、本研究では科学技術論文に対してアノテーションを行うアノテータの協力が不可欠であった。こうした文書の読解には高い専門性が必要となるため、さまざまな理工系分野の専門家にアノテーション作業を依頼するにあたって研究費が大変有効に活用できたと考えている。研究補助員は、データ作成のアノテーションツール開発の補助にあたり、本研究課題の効率的な推進において大きな役割を果たした。また、海外の研究者との緊密な連携により、細部の議論をより充実させることに繋がったほか、構築したデータセットを世界中の多くの研究者に活用してもらえるよう形式や配布方法を大幅に洗練させることが可能となった。

上記のように、数式グラウンディングの実用レベルでの自動化のためにはさらに継続した研究が必要であるが、本タスクは理工系の学術文献理解においてはさまざまな応用タスクの基盤となる重要な位置を占める。学術文献には未来のイノベーションに繋がる膨大な科学的知識が記述されているため、その解析技術の基盤開発の将来的な波及効果は極めて大きい。本研究の成果は、科学的知識の自動抽出、セマンティックな数式検索、論文から実行可能形式への自動変換などに応用でき、科学技術の研究・開発全体の効率化に寄与するだろう。また産業界とのつながりにおいては、ACT-X の枠組みで研究を進めていたことから民間企業からも一定の注目を浴びることとなった。実際、今後は民間企業においても過去に出版された膨大な書籍のOCRデータの解析・検索技術に関する研究を進める話に展開しており、中長期的にこうした研究に取り組む予定である。こうした書籍OCRデータの中には、もちろん数式を含むような文献もあるため、特に今回の数式グラウンディング研究を通して得られたベースライン技術や知見は、ここからの情報抽出や検索において大いに役立つものと期待される。本プロジェクトでの取り組みが、今後のより社会実装に近い研究開発プロジェクトに着実に繋がっていると考えられる。

(加速フェーズ実施後追記)

加速フェーズの成果により、数式識別子の文書内曖昧性という、データ資源の不足のために

これまでほとんど調査されてこなかった数式の重要な性質の一端が明らかになった。本研究では、データ資源の新規構築という NLP 研究のさまざまな段階の中でも特に経済的・時間的コストが大きい部分を、ACT-X の強力なサポート体制の下で、着実に進展させることに成功した。また、この曖昧性の解消に必要な情報種を分野に強く依存しない形で特定したことは、今後のデータ収集やモデル設計に向けて重要な指針となることが期待される。

加速フェーズでは特に、ACT-X 研究期間中に構築したデータセットを拡張した上で、構築したデータセットが実際に機械学習モデルの訓練と評価に利用可能な状態であることを具体的事例を以って示した。これにより本研究で検討したタスクやデータセットは今後の研究にスムーズにつなげることが可能となった。

5. 主な研究成果リスト

(1) 代表的な論文(原著論文)発表

研究期間累積件数: 2件

1. Takuto Asakura, Yusuke Miyao, Akiko Aizawa. Building Dataset for Grounding of Formulae – Annotating Coreference Relations Among Math Identifiers. In Proceedings of 13th Conference on Language Resources and Evaluation (LREC 2022). pp. 4851–4858, 2022.

学生アノテータとともに作成した数式グラウンディング用データセットに関する報告。15 本の arXiv 論文に現れる合計 12,352 個の数式識別子すべてに対してグラウンディング情報のアノテーションを行った。論文中の数式識別子の曖昧性は大きい、人手であれば高いアノテーション間一致率の下でグラウンディング情報をアノテーションできることを示した。この成果により、自動化への試みにおいて重要な ground truth データを確保することができた。

2. Takuto Asakura, Yusuke Miyao, Akiko Aizawa, Michael Kohlhase. MioGatto: A Math Identifier-oriented Grounding Annotation Tool. In 13th MathUI Workshop at 14th Conference on Intelligent Computer Mathematics (MathUI 2021).

数式グラウンディング用データセットを構築するために、申請者が独自に開発したアノテーションツールに関する報告。数式グラウンディングタスクの大きな特徴である「共参照情報」のアノテーションを効率的に行えるよう工夫が凝らされている。開発したツールはオープンソースソフトウェアとして公開している。

(2) 特許出願

なし

(3) その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

- 言語処理学会第 28 回年次大会 (NLP 2022) 委員特別賞 受賞
- NLP 若手の会第 16 回シンポジウム (YANS 2021) 口頭発表
- アノテーションツール MioGatto, <https://github.com/wtsnjp/MioGatto>
- グラウンディングデータセット,
<https://sigmathling.kwarc.info/resources/grounding-dataset/>