数学概念への数式グラウンディングのためのデータセット

朝倉卓人 [1, 4] André Greiner-Petter[2] 相澤彰子 [1, 3, 4] 宮尾祐介 [4]
総合研究大学院大学 [1] ヴッパータール大学 [2] 国立情報学研究所 [3] 東京大学 [4]
{asakura,aizawa}@nii.ac.jp andre.greiner-petter@t-online.de
yusuke@is.s.u-tokyo.ac.jp

1 はじめに

論文，専門書など理工系の学術文献（以下「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意義深い．科学技術文書には，他の文書と顕著に異なる特徴がいくつかあるが，数式の存在もそのような特徴の 1 つである．数式は様々な自然科学分野に共通のコミュニケーション手段であり，多くの科学技術文書において特に重要なアイデアを表現するために用いられている．そのため科学技術文書の解析にあたっては，文書中に現れる数式の意味や意図の推定が不可欠である．このように数式は重要であるにも関わらず，これまで自然言語テキスト中の数式は，一般に非言語的な要素とみなされ，積極的に研究されてこなかった．しかし実際には，自然言語の中に現れる数式には数々の言語的な現象が見られる．そのため数式に対しても計算言語学や自然言語処理に似たアプローチをもって理解を深めないことには，科学技術文書を深く解析することはできない．すなわち数学言語学 (mathematical linguistics) 研究および数学言語処理 (mathematical language processing) 技術の確立が必要不可欠である [3].

本稿では，こうした数学言語学や数学言語処理の研究に資するため，数式グラウンディングという新しいタスクと，その実現のためのデータセットを提案する．科学技術文書の読者が数式を理解するにあたっては，まず数式の**単語**境界を認識し，続いて数式内の各単語をそれぞれの意味する**数学概念**と紐付ける必要がある（図 1）．この 2 つのステップを**数学概念への数式グラウンディング**と呼ぶ．かかるグラウンディングは数学言語処理の最初のステップとなるべき過程だが，数式内では変数の曖昧性などの言語現象が見られるため，既に自明な操作ではない．例えば図 1 では，文章中の 1 つ目と 3 つ目の $\mathbf{y}$ は関数だが，2 つ目は関数ではなくベクトルであり，数式トークン $\mathbf{y}$ が複数の意味で用いられている．こうした曖昧性を乗り越えて計算機によるグラウンディングを実現するため，我々は 1 本の長編論文について，数式内に現れるすべての**識別子** (identifier) に対してグラウンディングに必要な情報を手作業でアノテーションしたデータセットを作成した．我々の言語リソースでは，直接各識別子にそれらと紐付けられるべき数学概念をアノテーションするのではなく，参照する数学概念ごとにクラスタを定義するようにした．つまり，図 1 の例では 1 つ目と 3 つ目の y は同一の関数オブジェクトを参照するクラスタに，2 つ目は別のオブジェクト（ベクトル）を参照するクラスタに紐付けられる．これにより，このデータセットの評価と応用がしやすくなっている．本データセットの最大の新規性は，1 本の論文に現れるすべての識別子に対して，それぞれが指し示す数学概念を首尾一貫してアノテーションしたことである．また本研究により，単一の文書においても，1 種類の識別子が複数の数学概念を参照するということが，実際に行なわれていることが明らかになった．

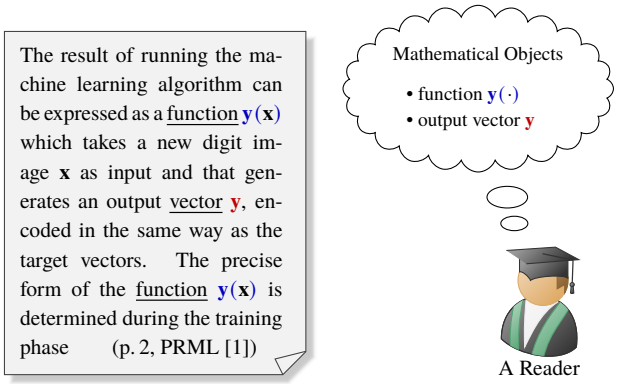

図 1 数式グラウンディング

2 関連研究

数こそ多くないが，これまでにも数学言語学や数学言語処理を指向する言語リソースはいくつか作成されてきた．arXMLiv [6] は XML 形式に変換された科学論文（数式は MathML 形式）から成る大規模なコーパスで，我々がアノテーションした論文データもこのデータセットから入手した．arXMLiv に含まれる XML に対して手作業でアノテーションが付与されたものとしては NTCIR-10 の数式情報抽出タスク向けのテストデータ [4] がある．NTCIR-10 データセットでは，数式の各要素に対して**説明**がアノテーションされている．例えば，論文中に現れる $\log(x)$ という数式に対しては「値 x について自然対数を計算する関数」といった説

明が与えられている．このデータセットの目的は我々のそれとよく似ているが，我々は各要素に対して，自然言語による説明のみならず接辞やクラスタの情報など，いくつかの付加情報を与えた．また，我々はより長い 1 本の論文に対して，一貫したアノテーションを行った点にも違いがある．

数学言語処理の応用としては，変数への型付けタスクが提案されている [7]．このタスクは科学技術文書中に現れる変数について**数学的な型**を与えることを目指す．例えば，次の文では変数 P には “parabolic subgroup”，N には “unipotent radical” がそれぞれ数学的な型として割り当てられる：

> Let P be a parabolic subgroup of GL(n) with Levi decomposition $P = MN$, where N is the unipotent radical. [7]

この型付けタスクは我々のいう数式グラウンディングと似ているが，少なくとも次の 2 つの点で異なっている．第一に数学的な型は，**数学概念**の一部で，下位の概念にあたる．第二に，Stathopoulos らの研究では自然言語中に単一のトークンとして現れる変数しかアノテーションの対象としていないが，我々のグラウンディングではより複雑な数式に含まれる識別子もすべて対象とすることを目指している．

また，別の応用として Part-of-Math タグ付けというタスクもある [8]．これは自然言語における Part-of-Speech タグ付けの数式版に相当する．Youssef の開発しているタグ付け器では，複数回のスキャンによりタグ付けを実行することが計画されており，現時点ではその第一スキャンの実装ができている．第一スキャンでは，添字・関数・開始デリミタなど，各トークンの数式内での役割をタグ付けする．このタグ付け器の実装には，約 2,800 の記号について，各分野で典型的な役割やカテゴリをまとめたデータベースが含まれている．現時点では計画段階にある Part-of-Math タグ付けの第二・第三スキャンでは各種の曖昧性解消や，NLP 技術を活用しつつ数式周辺のテキストを解析することによる，数式からのみでは収集できないセマンティック情報の抽出・付与が計画されている．我々のデータセットは，こうした機能拡張を行う際に役立つだろう．

3 数式グラウンディング

科学技術文書中の数式を理解するためには (1) 数式内での単語境界を認識し (2) 各単語をその意味する数学概念と結びつけるという 2 ステップ，すなわち数学概念への数式グラウンディングが必要となる．なぜなら各単語が参照している数学概念を把握しないことには，数式について正確に構文解析を行うことすらできないからである．

数式グラウンディングをより形式的に定義するため，ここでいくつか数式に対しても形態論の用語を導入しておく：

- **形態素（トークン）**は数式内で意味のある最小の単位である．プレゼンテーション MathML では 1 つの要素（タグ）がこれに対応する．形態素は 1 個の文字または記号（$x, \theta, \prime, \times, =, \sum$ など）の場合と，短い文字列（log や argmax）の場合がある．
- **単語**は単独で 1 つの数学概念を参照することのできる形態素のグループである．数式においては，単語は 1 つまたは数個の形態素から構成される．例えば x, x_i, $\stackrel{\text{def}}{=}$, log(·) は単語である．すべての単語には，その単語の核となる意味をもつ**語基**となる形態素が含まれ，場合により補助的な役割を担う 1 つ以上の**接辞**が含まれる．例えば単語 x_i において x は語基であり i は接尾辞（接辞の一種）である．
- **数学概念**は単語の意味である．数学概念は数学的な定義と，数式に対する解釈の一部を反映するものである．1 つの数学概念は 1 つの説明に置き換えられる．数学概念にはデータ型や，変数ならば値の範囲，関数ならばアリティ（引数の数）などの情報も含まれる．例えば π に紐づく典型的な数学概念は「円周率．その値は約 3.14」となる．数式において，すべての単語は 1 つの数学概念を参照する．

現時点では，数学概念の定義は「説明といくつかの付加情報から成るもの」と曖昧なままに留めている．これは，数学概念を厳密に定義する上で，どのような属性があれば十分なのかが未だ判然としないためである．とはいえ，現状の定義は，1 つの文書の数式に現れるすべての単語について，同じ数学概念を参照するもの同士をクラスタリングするのに十分である．数学概念の形式的な定義は，大きなデータセットを作成した上で検討する．

4 提案データセット

データセットの作成にあたって使用した論文データは arXMLiv:08.2018 [2] から入手した．このデータセットには約 120 万本の論文データが含まれているが，その中から手作業でアノテーションを行うため，筆者らにとって読みやすい分野で，数十ページ程度の長さがあり，数式が十分に含まれているもの，ということを基準に対象の論文を選定した．そして，最終的に論文 *A Very Brief Introduction to Machine Learning With Applications to Communication Systems*[1] [5] を対象として選び，その論文に含まれるすべての数式について，アノテーションを行った．この論文は約 1 万単語 10 ページで，およそ 350 の数式を含んでいた．

今回は初めての試みであったので，対象を数式のすべての形態素とするのではなく，識別子に限定した．識別子は数式形態素の 1 種で，数式内で変数・関数・定数などを表す．識別子は一般に単一の文字（x, y, θ など）または短い文字列（sin や log）である．プレゼンテーション MathML において

1) https://arxiv.org/abs/1808.02342

は，識別子は必ず <mi> タグ（mi は “math identifier” を意味する）によってマークアップされるため，この基準は明確である．最初のアノテーション対象として識別子を選んだのは，一般的な数式においては識別子が主要な要素だからである．

アノテーションを行う前に，対象の XHTML データに対して前処理を行う．まず，本来 <mi> タグでマークアップされるべき識別子が，誤って別のタグでマークアップされているような箇所をルールベースの置換によって修正する．続いて論文中の MathML 数式を走査して (1) 識別子の種類をまとめた**辞書**のテンプレートと (2) 文書中における識別子の出現をまとめたリストをスクリプトにより生成する．このうち (1) の辞書は後に手作業で各識別子の意味（数学概念）を記入していくもので，いわば一部の書籍に付録としてある「記号の用法」リストを詳細にしたようなものである．なお識別子においては大文字・小文字および書体の違いは意味の違いを反映しており重要であることから，辞書においてもそれぞれ別の項目として扱う．

実際の手作業によるアノテーションは，これら 2 種類のリストを編集する形で行う．アノテータは論文を読みながら，各識別子の意味が定義されたり，新しい用法で用いられたりするたびに，自動生成した (1) の辞書テンプレートに数学概念を記入していき，辞書を作成する．さらに，識別子が現れるたびに (2) のリストに，各文脈においてその識別子が指し示す数学概念に対応する (1) の辞書項目へのポインタを記入していく．このリスト (2) への該当辞書項目記入は，我々がこのタスクのために開発した GUI ソフトウェアにより，効率的に作業を行う．

5 データセットの評価と解析

提案データセットが高品質で再現性のあるものであることを示すため，3 人のアノテータが独立にアノテーション作業を行った．まずアノテータ 1 は辞書作成と各出現に対する辞書項目（数学概念クラスタ）の紐付けのすべてを行った．その上で，アノテータ 1 が作成した辞書がアノテータ 2, 3 に渡され，彼らは独立に後半の作業（各識別子の出現と辞書項目の紐付け）を行った．なおアノテータ 2 は共著者であり，アノテータ 3 は機械学習の見識はあるが共著者ではない人物である．対象の論文では識別子の出現が 937 回あり，そのすべてを辞書項目と紐付ける作業にはおよそ 1 日を要した．

今回のデータセット作成におけるアノテータ間一致率を表 1 に示す．937 回の識別子出現のうち，14.09%にあたる 132 回の出現は，辞書項目を 1 つしか持たない（つまり，文書を通して単一の意味で用いられている）識別子のものであった．このように選択肢が 1 つしかないものについてはアノテータ間で一致することは最初から明らかであるが，表に示した一致率にはこうした自明な一致も含まれている．1

表 1 アノテータ間一致率

	一致率	接辞ミスマッチ
アノテータ 2	904/937 (96.48%)	2/33 (6.06%)
アノテータ 3	824/937 (87.94%)	60/113 (53.10%)

種類の識別子の辞書項目のうち，アノテータ間で接辞タイプが異なる辞書項目が選ばれた場合，これは「単語境界の識別」の時点で見解の相違が生じていることを意味する．例えば識別子 p について，1 人のアノテータが単語 $p(\cdot \mid \cdot, \cdot)$（パラメタ化された条件付き確率）に含まれる形態素とアノテーションしたのに対し，別のアノテータが単語 $p(\cdot \mid \cdot)$（単なる条件付き確率）としてアノテーションしたような場合がこれに該当する．このような不一致は注目に値するため，すべての不一致（分母）の中で接辞が合わない不一致（分子）の割合を表 1 の最右列に記載している．

不一致の中身を分析すると，アノテーションの不一致にはカスケード効果が認められた．これは 1 つの識別子の出現で参照される数学概念についての見解が異なる場合，その出現と関連する後続の出現についても見解が分かれる結果となるためと考えられる．我々のアノテーションにおいて，$\mathcal{D}$ が参照する数学概念に関するアノテータ間での見解の相違が最も本質的であった．対象の論文では $\mathcal{D}$ は各種の学習タスクの訓練データを表すのに用いられているが，訓練データへの仮定（例えば，各データポイントが何らかの確率分布にしたがうか否か）は節により異なる．そして，節によっては，そこで扱われている訓練データへの仮定が明確に述べられておらず，アノテーションの不一致につながった．

> Under this assumption, the data set $\mathcal{D}$ is not necessary, . . . [7]

アノテータ 1, 2 間の不一致は，単純なミス（接辞ミスマッチの 2 つ）を除いてはほとんどが $\mathcal{D}$ の参照する数学概念の解釈相違に起因するものであった．アノテータ 1, 3 間にはより多くの不一致があったが，同じパターン（同じ識別子に対してアノテータ 1, 3 が各々同様に obj_1, obj_3 をアノテーションしている状況）が頻出しており，113 の不一致は 40 パターンに分類できた．

図 2 は識別子の各出現と，今回のアノテーション結果を一覧できるようプロットしたものである．x 軸が文書中の位置を表しており，左端が文書の開始地点で，右に行くにしたがって節が進んでいく．y 軸はアノテーションされた辞書項目を表し，左側に各辞書項目が用いられた回数のヒストグラムを表示している．この図から，文書中の数式で参照されている数学概念の分布には偏りがあることがわかる．科学技術文書の数式においても，厳密ではないが**スコープ**の存在が認められる．例えば，識別子 x のスコープは §3.2, 3.5, 3.6, 5.1 の開始位置で切り替わっている．しかし，散発的にスコープ外の数学概念を参照する識別子の出現も見られ，特に t や p には x ほど明確なスコープは認められない．

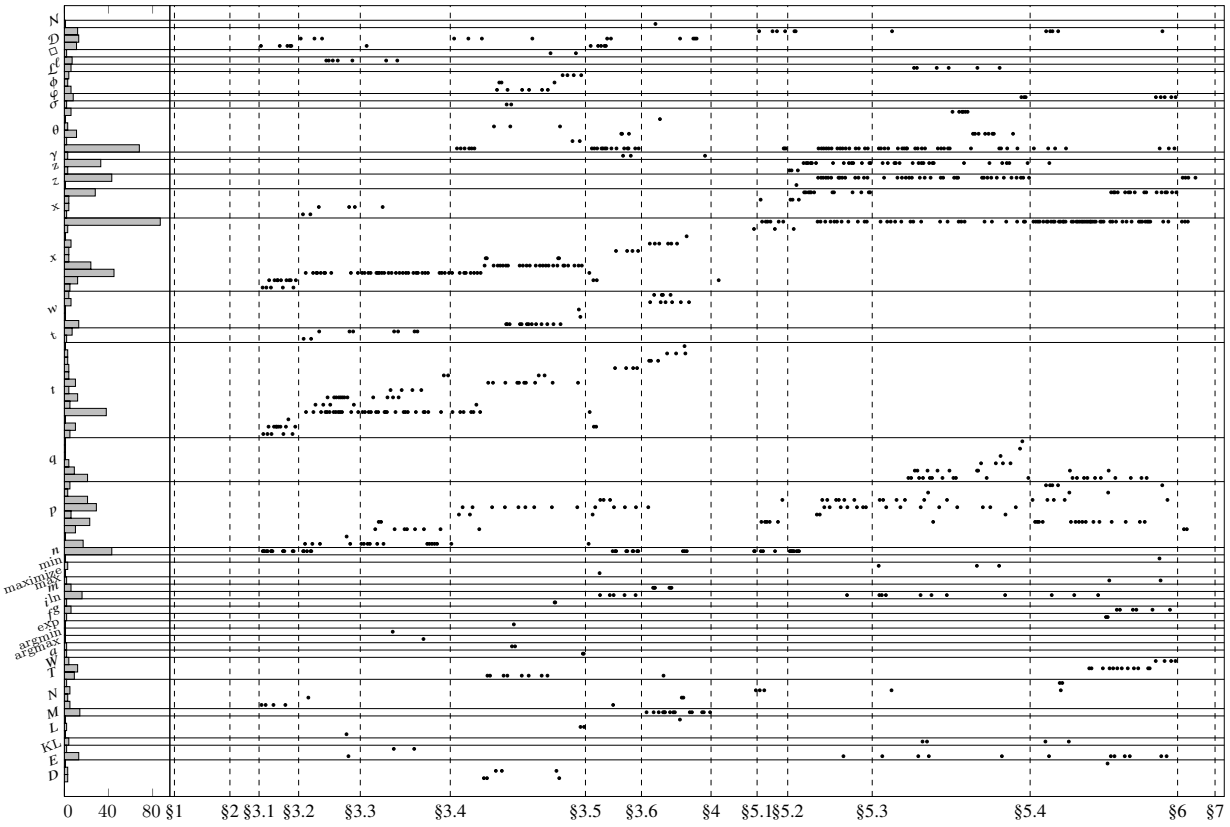

図 2　対象論文における形態素と数学概念の分布

興味深い個別事例として，アノテーションの対象論文には識別子の書体の使用法に関して言及することで，複数の識別子の参照する数学概念を一括的に指示する文があった．

> Throughout, we use Roman font to denote random variables and the corresponding letter in regular font for realizations. [7]

アノテータは例えば x と x の違いを認識するためには，上記に注意を払う必要があった．

6　結論と今後の展望

本研究では，高いアノテータ間一致率で長編の論文に一貫した数学概念クラスタのアノテーションを行うことができた．また，我々の分析結果により，実際に単一の科学技術文書において，識別子には不明確ながらスコープのようなものが見られることがわかった．今後は本データセットをさらに多角的に拡張し，数学言語処理技術のさらなる発展に寄与していきたい．

謝辞　本研究は JST CREST (JPMJCR1513) の支援を受けて行なわれたものである．

参考文献

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] D. Ginev. *arXMLiv:08.2018 dataset, an HTML5 conversion of arXiv.org*. SIGMathLing—Special Interest Group on Math Linguistics. 2018. URL: https://sigmathling.kwarc.info/resources/arxmliv/.

[3] M. Kohlhase and M. Iancu. “Co-Representing Structure and Meaning of Mathematical Documents”. In: *Sprache und Datenverarbeitung, International Journal for Language Data Processing* 38.2 (2014), pp. 49–80.

[4] G. Y. Kristianto, G. Topić, and A. Aizawa. “Utilizing dependency relationships between math expressions in math IR”. In: *Information Retrieval Journal* 20.2 (2017), pp. 132–167.

[5] O. Simeone. “A Very Brief Introduction to Machine Learning with Applications to Communication Systems”. In: *IEEE Transactions on Cognitive Communications and Networking* 4.4 (2018), pp. 648–664.

[6] H. Stamerjohanns et al. “Transforming Large Collections of Scientific Publications to XML”. In: *Mathematics in Computer Science* 3.3 (2010), pp. 299–307.

[7] Y. Stathopoulos et al. “Variable typing: Assigning meaning to variables in mathematical text”. In: *NAACL2018*, pp. 303–312.

[8] A. Youssef. “Part-of-Math Tagging and Applications”. In: *CICM2017*, pp. 356–374.