数理・情報のフロンティア
2020 年度採択研究者

2021 年度
年次報告書

朝倉　卓人

東京大学 大学院情報理工学系研究科
大学院生（博士課程）

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

§1. 研究成果の概要

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」というタスクの自動化を目指している。この目標の達成に向けて、今年度は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出来上がった。

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複数発見することができた。数式を含む文書をコンピュータによって処理しやすい形（形式表現）に変換する上で必要な、数式トークンの導入のされ方の区別を自動的に行うため、追加のデータセット構築を行う準備も進んでいる。

ここ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自然言語処理分野のトップ国際会議の 1 つである LREC 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。開発した専用ソフトウェア 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。

【代表的な原著論文情報】

1) 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).
2) Takuto Asakura, Yusuke Miyao, Akiko Aizawa. Building Dataset for Grounding of Formulae – Annotating Coreference Relations Among Math Identifiers. In 13th Conference on Language Resources and Evaluation (LREC 2022).
3) 朝倉卓人, 宮尾祐介, 相澤彰子. MioGatto による数式グラウンディングデータセットの構築. 言語処理学会第 28 回年次大会 (NLP 2022).