研究課題別事後評価結果

１．研究課題名：　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

２．個人研究者名

朝倉　卓人（東京大学大学院情報理工学系研究科　大学院生）

３．事後評価結果

本研究では、数式グラウンディング、すなわち数学論文中に記述された数式記号（例えば、x や y）がどのような数学概念に対応しているかを同定するという、極めて挑戦的な課題に取り組んだ。その課題の第一歩として、15 編の論文について、その中の記号（総数 1.2 万）とその数学概念を対応付けたデータセットを構築した。その結果、本来曖昧性のない明確な記述が必要とされる論文において、多くの記号が 2 つ以上の解釈が可能（例えば論文の途中で意味が変わる）という、驚くべき結果を定量的に得ている。現時点では、こうした課題の困難性の把握、データセット構築環境(Mio Gatto)、およびデータセット記述フォーマットを、海外の研究者とのディスカッションも行いながら実現した段階である。今後はこれら知見を活かし、自動対応付け（アノテーション）や、意味の切り替わり（スコープ）の自動判定を実施予定とのことである。同技術は、知の集積としての論文に記述された内容を二次利用するためには不可避であり、その嚆矢として、本研究の今後の発展が期待される。

（加速フェーズ）

上記の評価を受けて研究実施期間を 1 年間延長し、加速フェーズを実施した。

加速フェーズでは、数式グラウンディングの精度向上を目的として、データセットの大規模化および機械学習による曖昧性解消モデルの実装と検証を実施した。データセットについては、倍増以上の 40 編の論文を収録するに至った。また、曖昧性解消モデルについては、人間のアノテーターでの精度が 90%程度であるのに対し、それに比肩する 85%を達成するという成果を得ている。さらにそのモデルが数式に関するどのような特徴を重視しているかについても検証しており、特に数式の位置と接辞タイプ（数式周辺のローカルな構造）が重要と言う興味深い結果も得ている。さらに、特定分野（具体的には自然言語処理）の論文のみを用いた学習実験を介して、この特徴量の有効性が、分野を越えた普遍的なものであることも確認している。こうした成果は、数式を利用した学術論文の二次利用において極めて重要な一歩であると考えられ、今回得られたデータセットや結果を端緒として、今後世界を巻き込んだオープンな技術流を生むことを期待する。