数理・情報のフロンティア
2020 年度採択研究代表者

2020 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

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を目指している。

今年度の研究では、こうした曖昧性の解消を自動的に行うことができるような技術を開発するために、まず人の手によって曖昧性を解消した場合の結果と、その解消に必要であったテキスト中の情報を、実際の科学技術文書に対して注釈付け(アノテーション)した。このようなアノテーションは初めての試みであるため、アノテーションに必要なグラフィカル・ユーザ・インターフェース (GUI) を持つ専用のアプリケーションを独自に開発した。この専用アプリケーションを用い、高度な専門知識を持つ作業者(アノテータ)が、機械学習分野の論文に対して必要な情報のアノテーションを行った。現在までに、4 名のアノテータにより同一文書に対するアノテーションを行ってきたところ、先述の曖昧性は人間であれば概ね 90%以上の精度を持って正しく解消できることがわかった。また、曖昧性の解消に必要となったテキスト中の情報も同様にアノテーションし、該当するテキスト断片を 200 以上収集することができた。

今年度開発したアノテーション用の専用アプリケーションおよびデータセットは、公開の準備が整い次第、次年度以降に順次公開していく予定である。