

研究課題別事後評価結果

1. 研究課題名： 構造理解に基づく大規模文献情報からの知識発見
2. 研究代表者名及び主たる研究参加者名（研究機関名・職名は研究参加期間終了時点）

研究代表者

松本 裕治（理化学研究所革新知能統合研究センター チームリーダー）

主たる共同研究者

佐藤 健（国立情報学研究所情報学プリンシプル研究系 教授）

乾 健太郎（東北大学大学院情報科学研究科 教授）

相澤 彰子（国立情報学研究所コンテンツ科学研究系 教授）

鶴岡 慶雅（東京大学大学院情報理工学系研究科 教授）

森 純一郎（東京大学大学院情報理工学系研究科 准教授）

狩野 芳伸（静岡大学大学院情報学領域 准教授）

進藤 裕之（奈良先端科学技術大学院大学先端科学技術研究科 助教）

3. 事後評価結果

○評点：

A+ 非常に優れている

○総合評価コメント：

学会や出版社、更にはウェブから得られた論文のPDFコピーを含む大規模文献データベースを対象に、①個々の論文PDFを機械可読に変換し、NLPの適用を可能にした。論文PDF中の表やグラフも機械可読にし、数値の利用も可能にした。②機械可読になった論文に、NLPを用いて、アノテーション支援システムを開発して、抽出すべき概念や関係に人手でアノテーションをつけると共に、この作業を学習して可能な限り自動化する機械学習機能を開発して、アノテーションの効率化を実現した。③個々の論文内における概念と関係のネットワーク表現をデータベース化し、その部分構造を変数付きで検索条件として指定することにより、条件を満たす部分ネットワークのインスタンスと、それを含む本文の文章を検索可能なシステムを開発した。④複数論文間の関係に関しても、概念と引用関係に基づく、論文間のネットワークの構築とその中での論文分類を行い、特定トピックスにおける研究の動向の可視化を可能にした。⑤概念とその関係に基づくネットワークを複数論文に跨って辿る論理推論機構の開発により、個々の論文で述べられている断片的知見を繋ぎ合わせることによって初めて推論できるような新しい知見の発見を支援する技術を構築した。⑥これらの成果を、物質材料物性や脳科学、法学分野の論文や文書資料に適用し、これらの分野での知識発見支援システムを開発した。国際法律文書処理コンペティションでは関連判例検索タスクにおいて2018年、2019年と連続優勝している。最終年度にはCOVID-19のパンデミック対応として、急速に増える膨大なCOVID-19に関するバイオメディカル文献にもこれを適用し、トランプ前大統領が感染予防のために用いた抗マラリア薬ヒドロキシクロロキンの効果の可能性を再発見している。基盤として新たに構築した英語の機能表現および動詞、形容詞に関連する複単語表現の辞書は、Linguistic Data Consortium (LDC)から公開している。

中間評価の時点では、①～④に関しては順調な進捗状況であったが、⑤に関しては遅れており、この部分への注力を依頼した。その結果、⑤に関して特に最終年度に大きな進展が認められた。

科学技術における先端研究において、研究者の専門化が進む一方で、イノベーションには複数学問領域に跨る最先端知識を用いた仮説推論とその検証がますます必要となっており、膨大な文献からの知識発見技術の研究開発の重要性は一層高まっている。本プロジェクトは我が国におけるこの技術の研

究開発の中核となるメンバーにより遂行されており、目標に向けて着実に成果を積み上げてきたことは高く評価できる。