「科学的発見・社会的課題解決に向けた各分野のビッグデータ利活用推進のための次世代アプリケーション技術の創出・高度化」
平成 27 年度採択研究代表者

H28 年度
実績報告書

松本裕治

奈良先端科学技術大学院大学 情報科学研究科
教授

構造理解に基づく大規模文献情報からの知識発見

§1. 研究実施体制

(1)「G0」グループ
① 研究代表者:松本 裕治 (奈良先端科学技術大学院大学情報科学研究科、教授)
② 研究項目
・論文テキスト解析のための辞書および言語解析ツールの開発
・単語・表現・文の意味的類似度に関する研究
・論文アブストラクトの構造化に関する研究
・エンティティリンキングおよび関係抽出に関する研究

(2)「G1」グループ
① 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)
② 研究項目
・自然言語処理と事例ベース推論における類似度学習を融合した観点に基づく類似判例検索

(3)「G2」グループ
① 主たる共同研究者: 乾 健太郎(東北大学大学院情報科学研究科、教授)
② 研究項目
・仮説推論に基づく論述構造の解析

(4)「G3」グループ
① 主たる共同研究者: 相澤 彰子(国立情報学研究所コンテンツ科学研究系、教授)

② 研究項目

・文書構造の解析のための訓練用データの作成および性能評価、および、閲覧デモシステム上で予備的な評価

(5)「G4」グループ

① 主たる共同研究者: 鶴岡　慶雅(東京大学大学院工学系研究科、准教授)

②　研究項目

・論文の深い意味理解のための基盤技術の開発

・単語や文の意味表現技術の開発

・高精度関係抽出技術の開発

・高精度エンティティリンキング技術の開発

(6)「G5」グループ

① 主たる共同研究者: 森　純一郎 (東京大学政策ビジョン研究センター、准教授)

② 研究項目

・大規模引用ネットワークおよび文献テキストの構造的関係性に基づく潜在関連知識の抽出

・引用関係およびテキスト類似度に基づく論文ネットワーク分析

・異種多層ネットワークの表現学習

・異種多層ネットワークからの知識抽出

(7)「G6」グループ

① 主たる共同研究者: 狩野　芳伸(静岡大学学術院情報学領域、准教授)

② 研究項目

・脳科学論文のテキストマイニングと応用

§2. 研究実施の概要

科学技術論文などの専門性の高い文書を解析し、研究者や技術者の支援を目指すため、文書を柔軟に検索する方法と、重要な情報の抽出を行うための基盤技術やシステムの開発を目標としている。平成28年度は、そのための基盤技術として、文書構造の解析、文書内容の解析のための基本ツール構築と基本データの構築および構築支援ツールの開発と研究を重点的に行った。

PDF や画像フォーマットで与えられる論文ファイルを言語処理可能なテキスト形式に変換するためのツールを整備した。また、PDF 上に専門用語の範囲や用語間の関係を直接アノテーションできるツールを開発し、医学生物学分野や物質分野の数十本の論文に対して、重要な用語の出現箇所や用語間の関係をアノテーションしたデータを構築した。さらに、論文中の図表や数式の出現場所の特定と数式を解析する手法を開発した。また、論文中に出現する数式をその説明記述と対応付けて検索する数式検索手法を提案した [1)]。(G0, G3, G6 グループ)

言語の統語・意味解析について、単語間の修飾関係を高い精度で解析するための手法の提案 [2)]、複数の単語からなるまとまった表現(複単語表現)の辞書構築を行った。また、単語の意味表現を越えた句の意味表現を計算する新しい手法を提案した [3)]。文書中の専門用語を認識するツール、名詞句間の共参照関係や、用語間の意味関係を解析するための基本データの整理を行った。さらに、文書の深い理解のため、文書の論述構造の仕様設計を行い、文書理解のための知識獲得とその利用に関する基礎的な研究を行った。(G0, G2, G4 グループ)

文書間の類似性とそれを利用した文書検索に関する研究を行った。法律文書の類似性について、多次元尺度法とトピックモデリング(LDA)を組み合わせた文書類似検索のためのシステムを開発した。また、科学技術論文に対して、目的や手法などいくつかの視点における類似度を定義する手法を開発し、類似性に基づく論文検索インタフェースを試作した。また、論文内の専門用語や用語間の関係などを表示する論文閲覧システムを開発した。(G0, G1, G3 グループ)

科学技術論文データベースから効率よく論文情報や引用情報を利用する共通基盤の構築を行っている。論文フルテキストデータを取得するシステムのプロトタイプ実装を行った。また、代表的な文献データベースである Web of Science から文献の大規模な書誌情報ならび引用情報の収集を行い、分析用の高速なデータベースの構築を行った。他の主要な文献データベースである Scopus や PubMed などからも書誌・引用情報の収集も進め、プロジェクト全体において大規模文献情報を利活用できる基盤を構築している(G5, G6 グループ)。収集した情報を用いて、文献間の複数の大規模な引用ネットワークならびにテキスト類似ネットワークを構築した。(G6グループ)。

来年度は、グループ間のツールの統合をさらに進め、統合的なシステム構築を行う予定である。

1) Giovanni Yoko Kristianto, Goran Topic, and Akiko Aizawa, “Utilizing Dependency Relationships between Math Expressions in Math IR,” Information Retrieval Journal, 2017

2) Hiroki Ouchi, Kevin Duh, Hiroyuki Shindo, and Yuji Matsumoto, “Transition-Based Dependency Parsing Exploiting Supertags,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol.24, Issue 11, pp.2059-2068, 2016

3) Kazuma Hashimoto and Yoshimasa Tsuruoka, “Adaptive Joint Learning of Compositional and Non-Compositional Phrase Embeddings,” Proceedings of ACL, 2016, pp. 205-215.