「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

H27 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
・テキスト解析基盤技術および文書構造解析
・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

・言語・分野横断的な知識獲得を可能にする論文構造解析手法の研究

(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（東京大学大学院工学系研究科、特任講師）

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

科学技術論文等、専門性の高い文書を自動解析し、研究者や技術者の支援を目指すため、文書の様々なレベルでの意味的類似性を測ることが重要と考えている。平成27年度は、そのための基盤技術として、言語解析技術に関する研究と意味的類似性に関する研究を重点的に行った。

言語解析技術としては、英文の品詞解析や文分割、専門用語の意味同定、動詞と名詞の意味関係、名詞間の意味関係などの基盤技術開発を主にG0, G4グループを中心に行った。また、論文等で重要な情報として、本文以外の図や表、および、そのキャプションなどを正確に抽出することが挙げられる。G3グループでは、PDF 等のフォーマットで与えられる論文ファイルをユーザに提示される表示画面上での位置情報と対応をとりながら、言語処理可能なテキスト形式に変換するためのツール群の整備と機能拡張に取り組んだ。

意味的類似性に関しては、文脈を考慮しない独立した単語の意味関係だけでなく、文脈の情報を考慮した意味関係に関する研究[2]、単語から構成される句の意味が単語の意味から必ずしも推定できないような非構成的な意味を推定する技術の開発[3]、文全体の意味を単語の意味から計算する技術とさらにそれを単語の表現にフィードバックすることによって文の類似性をより正確に表現する技術の開発[1]などを行った(G0、G2、G4グループ)。

専門文書における論述構造や、裁判の判例文の構造的な類似性を測るための基礎的な調査および研究を開始した(G1、G2グループ)。言語基盤技術開発に関わる他のグループとの共同研究を今後すすめる予定である。

テキスト以外の情報に基づく論文間の関係解析に関する研究を行った。引用ネットワークの時系列変化ならびにダイナミックトピックモデルによる学術分野の構造変化の特定に関する基礎的な方法論を構築した。さらに、構築した大規模引用ネットワークに対して、それらを特徴量化するための表現学習手法の検討を行い、論文の引用ネットワーク構造から学習された表現ベクトルを用いて論文の将来のインパクトを予測する基礎的な方法論を構築した(G5グループ)。また、主要な出版社の論文フルテキスト提供方法ついて調査を進め、同時に、各社の指定する API 経由でフルテキスト情報を自動取得するクローラーのプロトタイプ実装を行った(G6グループ)。

[1] Masashi Tsubaki, Kevin Duh, Masashi Shimbo, Yuji Matsumoto, ″Non-Linear Similarity Learning for Compositionality,″ Proceeding of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), February 2016.

[2] 大野雅之, 井之上直也, 松林優一郎, 岡崎直観, 乾健太郎, “分散表現による文脈情報を用いた選択選好モデル,” 言語処理学会第 22 回年次大会, pp.885-888, March 2016.

[3] 橋本和真, 鶴岡慶雅, “構成性と非構成性を同時に考慮した動詞句の表現学習,” 言語処理学会第 22 回年次大会, pp.661-664, March 2016.