「科学的発見・社会的課題解決に向けた各分野のビック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

H29 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 (東京大学数理情報教育研究センター　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を開発している。平成29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と基本データの構築および論文検索インタフェースの開発と研究を重点的に行った。

PDF 論文の構造解析ツールの開発、および、アノテーションツールの開発を行った。開発したツールを自然言語処理分野の PDF 論文に適用して、引用情報や図表・数式画像とあわせて、当該分野の最先端の研究を網羅するコーパス資源として整備した(G0,G3)。論文等の専門文書に頻出する複雑な文の解析のための解析手法 [1]、および、並列構造の解析手法 [2]の提案を行い、高い解析性能を達成した。

文の意味解析の高度化を目指し、抽象意味表現(AMR)に基づく文の意味解析器を開発した(G1,G4)。論述構造を同定するため、論述文内の主張−根拠の関係、主張−反論の関係を自動解析するモデルを構築した。また、論文から概念間の意味関係を抽出するモデルを構築・評価した。さらに、昨年度までに設計した論述構造コーパスを用いて、論述構造を自動解析するモデルのプロトタイプを構築・評価した(G2)。法律文の前提と効果に関する記述を特定するための手法を検討し、深層学習に基づく手法の拡張を行った(G1)。

文書解析として、論文要約の研究を行い、論文の階層構造を利用した抽出型要約、および、文書分類とのマルチタスク学習を利用した抽出型要約 [3]の研究を行った(G4,G5)。

グループ間の共通データの構築を目指し、自然言語処理分野の論文データACL Anthologyデータの整備、Web of Science, Scopus や PubMed から大規模な書誌・引用情報の収集、論文フルテキストの自動取得クローラーの実現のため主要出版社とのTDM(Text and Data Mining)契約などを進めた(G3,G5,G6)。大規模文献情報を利活用できる基盤を構築し大規模な引用ネットワークの分析に基づいて学術領域の動向を抽出、可視化するための手法の研究開発を行った(G5)。

文書の類似性に関する研究とそれを利用した類似文書検索の研究を行った。論文データに対して、テキスト情報と引用情報を用い、目的や手法など複数の視点からの類似度計算の手法を提案した。一般的なドメインから法律ドメインへ適用に取り組み、ドメインの知識を組み込むことで文章の類似性のための深層学習の性能を向上させることにも成功した。これらを利用した文書検索のためのインタフェース構築を行った(G0,G1)。

1) Akihiko Kato, Hiroyuki Shindo and Yuji Matsumoto, English Multiword Expression-aware Dependency Parsing Including Named Entities, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Volume 2, pp.427-432, Vancouver, Canada, 2017.

2) Hiroki Teranishi, Hiroyuki Shindo, Yuji Matsumoto, Coordination Boundary Identification with Similarity and Replaceability, Proceedings of the Eighth International Joint Conference on Natural Language Processing, Vol.1, pp.264-272, Taipei, Taiwan, 2017.

3) Masaru Isonuma, Toru Fujino, Junichiro Mori, Yutaka Matsuo and Ichiro Sakata, “Extractive Summarization Using Multi-Task Learning with Document Classification,” Proceeding of Conference on Empirical Methods in Natural Language Processing, pp.2101-2110, 2017.