「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」
2015 年度採択研究代表者

2018 年度
実績報告書

松本　裕治

奈良先端科学技術大学院大学先端科学技術研究科
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. 研究成果の概要

G0 グループでは、論文等に頻出する並列表現を高精度で解析する言語解析法、および、代表的な複雑な言語表現である多単語表現解析のためのコーパスやコーパス管理ツールを開発した。また、PDF 文書の構造解析により、テキスト部分だけでなく、図表部分を識別する手法の開発を行った。目的・手法などの観点に基づく論文間類似性を定義し、論文構造を考慮した類似論文検索ツールのプロトタイプシステムを構築した。生命科学分野、物質科学分野の論文のエンティティ、および、エンティティ間の関係知識抽出法の高精度化に関する研究を進めるとともに、大規模な学習データが得られない状況での概念および関係抽出手法を提案した。

G1 グループでは、昨年度は主に、類似検索技術のうち、複数の長文からなる法律文書の類似検索手法について研究を行った。その基礎技術として、まず、複数の長文からなる法律文書の要約技術を深層学習を用いて開発し、その要約文同士を比較することで、法律文書そのものを比較する手法を凌駕する結果を得た。

G2 グループでは、計算機科学論文における知識抽出に関する技術的検討、およびプロトタイプシステムの開発を行った。具体的には、前年度までに研究開発した概念間関係抽出技術を拡張し、技術名を検索クエリとして、関連のある知識を有向グラフの形で可視化し、技術的な動向を俯瞰できるプロトタイプシステムを開発した。また、技術のメリット・デメリットに関する記述を自動抽出する計算モデルを構築・評価するために、92 本の論文からなる注釈付きコーパスを構築した。さらに、このコーパスに基づいて、自動抽出モデル及びプロトタイプシステムを構築し、技術的課題を分析した。発想支援の問題を知識グラフ補完の問題とみなし、構造化された知識ベースとテキスト情報の両方を用いた知識グラフ補完モデルについて、基礎的な検討に着手した。

G3 グループでは、PDF や XML 形式で流通する大量の学術論文を、もとの表示情報との対応をとりながら言語処理可能なテキスト形式に変換し、グループ全体で共有することを目指してい

る。前年度に引き続き、PDF 論文の文書構造解析のためのツールの開発およびデータセット構築に取り組んだ。自然言語処理分野の最先端の研究を網羅する論文アーカイブである ACL Anthology を対象として、網羅的なデータセットを構築してグループ内で共有するとともに、その一部を文コーパスとして一般に公開した。また、論文の内容を効率よく把握するための文書要約手法の研究に取り組み、文圧縮で高い性能を達成した。

G4 グループでは、 論文テキスト解析のための基盤的な言語処理技術の開発を主な目的として研究を行っている。当年度は、これまでに開発してきた文章解析や要約の技術を高度化することに加えて、新たに開発した主要な研究成果として、自然言語による質問応答（機械読解）技術、および効率的な強化学習による文章生成技術が得られた。

G5 グループでは、前年度までに構築した大規模な書誌情報ならび引用情報の収集基盤ならび分析用高速データベースを利活用し、論文、著者、組織、ジャーナルなど異なるエンティティから構成されるこれらの異種の多層ネットワークに対して、各エンティティを特徴量化するための表現学習手法の開発を行った。その上で、学習された各エンティティの表現を用いてエンティティとそれらの関係を含む知識ベースの設計と構築を進めた。具体的なデータセットとして ACL Anthology の文献データの関係性を可視化するとともに、対象論文群における成長領域の特定を行った。それらを機能として実装した大規模論文データの可視化システムを構築した。

G6 グループは、脳神経科学分野の論文を対象に当該分野の研究者に対して自動処理により有用な情報の提供を目指している。そのためには、論文フルテキストの自動取得、論文フルテキストの自然言語処理による解析、解析結果の可視化の三段階が必要となる。自動取得は実装ずみで運用中である。解析については、座標抽出の自動解析と評価、そのための新規アノテーションの付与と、アノテーションガイドラインの再整備を進めた。視覚化については、座標情報は 3 次元であるため、これを可視化するための 3 次元脳モデル表示系、さらに、これと連動した二次元スライス表示 UI を実装し、当該分野で必要とされる視覚化 UI のプロトタイプを構築した。

【代表的な原著論文】

1. Truong-Son Nguyen, Le-Minh Nguyen, Ken Satoh, Satoshi Tojo and Akira Shimazu: “Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts”, Artificial Intelligent and Law, Vol. 26, Issue 2, pp.169-199, 2018

2. Hiroki Ouchi, Hiroyuki Shindo, Yuji Matsumoto, “A Span Selection Model for Semantic Role Labeling”, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp.1630-1642, November 2018

3. Kimitaka Asatani, Junichiro Mori, Masanao Ochi, and Ichiro Sakata, “Detecting trends in academic research from a citation network using network representation learning”, PloS ONE, Vol. 13, No. 5, e0197260, 2018

§2. 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先端科学技術研究科　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　(東京大学大学院情報理工学系研究科　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