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

2019 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／理化学研究所革新知能統合研究センター
教授／チームリーダ

# 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、科学技術論文中の複雑な文の原因である並列表現を高精度で解析する言語解析法、および、複単語表現解析のためのコーパスやコーパス管理ツールの開発を引き続き行った。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、および、表やグラフの内容を読み取るための手法の開発を行った。生命科学分野、物質科学分野の論文中に記述されたエンティティの認識、および、エンティティ間の関係の分類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エンティティおよびエンティティ間の関係分類手法の高性能化を行った。

G1 グループでは、本年度は主に、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。COLIEE 2019 (the 6th Competition on Legal Information Extraction and Entailment) の関連判例検索タスクにおいて、判例を直接比較するのではなく、判例の要約同士を比較することで関連度の計算を精緻化した。ただし、COLIEE 2019 には学習用の要約がなかったため、COLIEE 2018 のタスクで用いた(判例・要約)のペアで学習させたモデルを用いても性能が保てることを示し、2 年連続優勝した。

G2 グループでは、論文集合中の別々の論文から別々に抽出した断片的知識を繋ぎ合わせてユーザに提供し、新たな知識の発見を支援することを目標としている。本年度は、データベース等の構造化知識ベースと百科事典等の非構造化知識ベースを統合的に用いて新たな知識の推論をすることが有効かどうかを検証した。その結果、①知識ベースを記号の形で保持しその上でマルチホップな推論を行う熟考的予測モデル、②知識とその推論結果を連続空間上にあらかじめ埋め込む即応的予測モデル、の二種類の推論方式について有望な実験結果が得られた。次年度では、推論効率と説明性の面で相補的な①と②のアプローチを統合し、推論効率のよい説明可能な推論機構を構築し、これを知識発見支援システムとして整備していく予定である。

G3 グループでは、科学技術論文解析の基本として必須の処理である 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する処理を実施した。それによって、言語や非言語情報の意味解析が可能になり、論文本文に記載されている詳細な内容を素早く検索したり把握したりすることが可能になる。この目標に向けて本年度は、複数文からなるパラグラフの内容を一文で簡潔に表すための要約手法の研究に取り組み、言い換えに基づく手法を新たに提案して有効性を示した。また、論文中の数学記号や図の意味を周辺テキストやキャプションから獲得して検索や理解に役立てるための手法の検討に取り組んだ。さらに、自然言語処理分野の網羅的な論文アーカイブである ACL Anthology に本研究で開発した文書構造解析ツールを適用し、データセットおよび論文閲覧のデモシステムの更新・公開を引き続き行った。

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。当該年度は、複数言語で共通するベクトル空間において適切な単語ベクトル表現を計算する技術、多様な文脈情報を利用して概念間の関係を予測する技術、および本グループと他グループで開発した関係抽出技術を統合するためのシステムに関して研究を行った。論文発表に関しては、前年度の研究成果が自然言語処理に関する主要な国際会議および論文**誌**

に採録された。

G5グループでは、本年度は昨年度から引き続き論文、研究者、ジャーナル、研究機関など異なるエンティティから構成される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研究開発を行った。また、大規模文献情報処理の要素技術として、教師なし学習による抽象型文書要約の手法についても研究開発を行った。これまでの研究成果を、大規模な引用ネットワークを分析可能な「学術産業技術俯瞰システム」に実装した。

G6 の研究の狙いは、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を提供することである。そのためには大きく分けて、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。本年度は、研究者が実際に必要とする機能を踏まえ策定したアノテーションガイドラインに従い、新規アノテーションの人手による付与を進めた。また、アノテーション自動付与のために PDF と HTML の両形式のフルテキストデータに対し、本文や表データの抽出を行うシステムの構築を進めた。

【代表的な原著論文】

1. Vu Tran, Minh Le Nguyen, and Ken Satoh, “Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model”, Proceedings of 17th International Conference on Artificial Intelligence and Law, 2019

2. Masaru Isonuma, Junichiro Mori, and Ichiro Sakata, “Unsupervised Neural Single-Document Summarization of Reviews via Learning Latent Discourse Structure and its Ranking”, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL2019), pp. 2142-2152, 2019

3. Van-Thuy Phi, Joan Santoso, Van-Hien Tran, Hiroyuki Shindo, Masashi Shimbo, and Yuji Matsumoto, “Distant Supervision for Relation Extraction via Piecewise Attention and Bag-Level Contextual Inference”, IEEE Access, Volume:7, Issue:1, pp.103570-103582, 2019

## §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