戦略的創造研究推進事業　CREST
研究領域「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」
研究課題「構造理解に基づく大規模文献情報からの知識発見」

研究終了報告書

研究期間　2015年10月～2021年3月

研究代表者：松本裕治
（国立研究開発法人理化学研究所革新知能統合研究センター、チームリーダ）

§1 研究実施の概要

(1)実施概要

G0 グループでは、生命科学および物質科学の研究者との協働により、科学技術論文から抽出すべき情報や知識の特定化、利用者の要望の具現化など、専門家が論文の検索や新情報の調査を行う場面を想定し、大規模文献からの知識発見支援のツール構築と実システムの開発を目指した。特に、科学技術論文の構造的な解析に必要な文書解析や情報抽出の基盤技術の研究開発、および、グループ間で共有して利用する文書解析ツールやユーザインタフェースの開発を行った。本グループおよび G3 グループで開発した専門用語認識、および、G1 グループの関係抽出技術を統合して、概念間関係に基づく COVID-19 関連論文の検索インタフェースの開発と一般公開を行った。

G1 グループでは法律文書の類似検索のための基礎技術を開発した。それらは、条文の条件部と結論部を分離する技術、トピックモデリングによる法律文書分類技術、深層学習を用いた文書要約技術、要約技術に基づいた類似文書検索技術、深層学習を用いた関係抽出である。関係抽出については、他のグループとの COVID 関連の論文における情報抽出と組み合わせることで論文の中から隠れた関係を発見する手法の開発へと展開した。

G2 グループでは、G3 グループの学術論文リソースを用いて、学術論文に記述された科学的知識(概念間の関係や属性など)を抽出する基礎技術を開発し、さらにこれを周辺の文脈(知識の根拠など)まで含めて深くマイニングする基礎技術を開発した。また、個々の学術論文から抽出した断片的知識と人手により精緻に整備された構造化知識ベースを統合的に用いて、複数の論文にまたがる未知の知識を推論する機構を構築した。また、これを応用し、COVID-19 に関する学術論文および生物医学分野の構造化知識ベース UMLS を統合的に解析し未知の知識の発見支援に繋げるシステムを G4 グループと連携しながら開発し、一般公開した。

G3 グループでは、論文の表示構造、論理構造、意味構造を統合する論文解析手法の研究に取り組み、その成果をツールおよび言語資源の形で公開して、他のグループとの連携をはかった。PDF 形式で流通する大量の学術論文の構造を解析して、言語処理可能なテキスト形式に変換することは、学術論文からの知識獲得のために必須の処理である。特に、PDF 上での表示位置との対応を維持しながら論文の論理構造を解析することで、論文本文に記載されている情報の閲覧支援や、論文中の非言語情報の意味解釈が可能になる。また、自然言語処理の研究分野に焦点を絞って、分野の網羅的な論文アーカイブの解析結果に基づくオンライン閲覧デモシステムを構築・公開した。

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、単語や句のベクトル表現をコーパスから計算する技術、論文の章構造を利用した抽出型要約技術、長期記憶を持つリカレントニューラルネットワーク、自然言語による質問応答技術、効率的な強化学習による文章生成モデル、教師なし学習による事前学習モデルを利用した関係抽出モデル等の研究を行い、自然言語処理に関する難関国際会議・論文誌(ACL 2016, EMNLP 2017, NAACL-HLT 2019, Computational Linguistics 2019, ACL 2020)等でこれらの研究成果を発表した。また、G2 グループによる知識ベースとテキスト情報の両方に基づく既存の知識補完モデルと、本グループで開発した関係抽出システムを統合し、単一のユーザインタフェースで概念間関係検索が可能なシステムを G2 グループと連携して開発した。

G5 グループでは、代表的な文献データベースから大規模な書誌情報ならび引用情報の収集を行い、引用ネットワークを構築し、プロジェクト全体において大規模文献情報を利活用可能な基盤を構築した。これらの基盤を元に、大規模文献情報をインタラクティブに分析可能なウェブシステムである「学術産業技術俯瞰システム」を開発した。その上で、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い、学術領域の動向や領域の融合・離散を可視化できることを明らかにした。また、他グループとの連携により COVID-19 に関連する文献の引用ネットワークを解析し、科学的エビデンスや重要技術などの情報を抽出し、その解析結

果を広く一般に公開した。また文書要約技術に関する研究にも取り組み、それらの成果を複数の主要な国際学会で発表を行うとともに産業応用を行った。

G6 グループでは、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を目的に、脳科学論文のテキストマイニングを実行するためのコーパス、リソース及びツールの構築を行った。まず脳科学分野の研究者からテキストマイニングのニーズを探り、技術的可能性を検討したうえで、G0 および G7 グループの成果も利用し実現に必要なテキストマイニングツールとそのために必要なコーパスおよびリソースを構築した。これらリソースを用いて、アノテーションの付与されていない論文についても座標抽出を含む自動アノテーション付与を行えるツールを実装した。さらに、論文本文とアノテーションの統合検索システムを実装し、3 次元・2 次元脳座標 UI と統合したウェブアプリケーションを構築した。

G7 グループは、2020年度に G0 グループから分岐し、G0 グループで行っていた科学技術文書処理の基盤技術として、PDF 形式の論文中の本文、図、表の認識と抽出および、図（主として折れ線グラフ）の読み取り、表内の項目の認識の手法開発を担当した。同処理のツール化を行い、他グループ（G3 および G6 グループ）へ提供した。また、材料分野の重要な情報として材料の合成プロセス情報抽出に関する研究を実施した。

（2）顕著な成果

＜優れた基礎研究としての成果＞

1．構造化知識とテキスト知識の統合的仮説推論機構

概要：テキスト集合から抽出した断片的知識と人手により精緻に整備された構造化知識を統合的に用いて新たな知識を推論し、これを推論根拠（説明）とともに提示できるようにする仮説推論機構について先駆的な研究を実施、NLP の国際会議 PACLIC2019、EACL2021 で発表し、研究員が一連の研究を博士論文としてまとめた。また、テキスト知識に基づく仮説推論として、質問応答データセットに推論根拠を効率的かつ高品質に付与する手法を考案し、世界最大規模の推論根拠付き質問応答データセットを整備・公開した。この成果はNLPトップ国際会議 ACL2020 に採択され、言語処理学会第 26 回年次大会では言語資源賞を受賞している。

2．複雑な構造をもった文を高精度で解析する手法の開発

概要：科学技術論文に頻出する複雑な文構造の要因となっている複文構造と並列構造を解析するためのリソースと解析アルゴリズムに関する研究を行った。前者については、英語の機能表現および動詞、形容詞に関連する複単語表現の辞書を構築し、Linguistic Data Consortium (LDC)から公開するとともに、複単語表現を考慮した統語構造解析を行い、高い解析性能を達成した。この成果に対して 2019 年度言語処理学会最優秀論文賞を受賞した。後者については、並列構造の要素が互いに類似していることと置換可能であることを考慮した解析手法を提案し、最高の解析精度を達成した。この成果は、NLP トップ会議の一つである NAACL2019 にて成果発表を行った。

3．関連判例検索手法の提案

概要：深層学習を用いた文書要約技術を用いて、長文である法律文書から要約を取り出してそれを使って類似文書検索を行うことで、国際法律文書処理コンペティション（Competition on Legal Information Extraction and Entailment(COLIEE)）における関連判例検索タスクにおいて 2018 年、2019 年と連続優勝した。さらに、BERT を法律用に特化することで 2020 年の COLIEE において含意タスク部門で優勝した。また、これらと並行して行っていた関係抽出手法に基づき COVID-19 論文横断検索システムを構築した。この結果は、The 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL2021)にてデモ論文として採択された。

＜科学技術イノベーションに大きく寄与する成果＞

1．学術産業技術俯瞰システム
特許：萌芽論文予測システム、特願 2016-062744、特開 2017-174357、出願人 国立大学法人東京大学、発明者 坂田一郎 、森純一郎 、佐々木一 、原忠義
概要：大規模な書誌・引用分析を行うためのウェブシステム「学術産業技術俯瞰システム」を構築した。その上で、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い、引用ネットワークから学習したネットワーク特徴量を用いることで、学術領域の動向や領域の融合・離散を可視化できることを明らかにした。これらの研究成果は政府機関や企業の研究開発戦略立案支援に利活用されている。

2．脳神経科学分野のための高度な学術文献情報抽出検索システムの構築
概要：脳神経科学分野特有の脳座標情報を軸にした情報抽出システムと、脳マップとしての 2D/3D ユーザインタフェースを備えた情報検索システムを、脳科学分野の研究者の協力を得て実装した。脳科学研究の強力なツールとなりえる。

3．PDF 論文閲覧システム SideNoter
概要：PDF 論文の解析結果に基づく論文閲覧システム。セクション単位の検索に基づく横断的な関連情報レコメンデーション機能や、閲覧時にサイドノートに参考情報を自動表示する注釈機能などを備える。研究者、特に、学生等の初学者、他分野の研究者のための論文読解支援として有効。

＜代表的な論文＞

1．Kazuma Hashimoto and Yoshimasa Tsuruoka. Adaptive Joint Learning of Compositional and Non-Compositional Phrase Embeddings. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), pp. 205-215.
概要：論文テキスト中の各文の意味を解析する際に重要な役割を果たす、句の意味表現を計算する手法を提案した。この手法は、句の意味を計算する際に、その構成要素の単語の意味からボトムアップに計算するモデルと、句をそのまま利用して非構成的にその意味を計算するモデルを動的に組み合わせて最適化する手法である。イディオム表現の認識や、動詞句の類似度判定において世界最高精度を達成した。手法は汎用的であり、関係抽出や文分類など、さまざまなタスクに利用可能である。

2. Masaru Isonuma, Junichiro Mori and Ichiro Sakata, ″Unsupervised Neural Single-Document Summarization of Reviews via Learning Latent Discourse Structure and its Ranking″, In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL2019), pp. 2142-2152, 2019
概要：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手法を新たに開発した。提案手法は、文書の各文の間の親子関係を潜在的な談話木として推定し、木の各ノードに対応する潜在変数から要約を生成するエンドツーエンドの新たなモデルである。レビュー文書要約タスクにおいて、提案手法は従来の教師なし学習モデルを上回る精度を達成することを明らかにした。

3．加藤明彦，進藤裕之，松本裕治，複単語表現を考慮した依存構造コーパスの構築と解析，自然言語処理，Vol.26 No.4，pp.663-688，2019.
概要：複単語表現は統語的または意味的な非構成性を有する複数の単語からなるまとまりであ

る。統語的な依存構造の情報を利用し、かつ意味理解が必要なタスクでは、単語ベースの依存構造よりも複単語表現を考慮した依存構造の方が好ましい。本論文では Ontonotes コーパスに対して新たに形容詞複単語表現を注釈し，複合機能語と形容詞表現の双方を考慮した依存構造コーパスを構築した。さらに、動詞複単語表現も考慮した依存構造解析手法をいくつか提案し、その評価を行った。本論文は、2019 年度言語処理学会最優秀論文賞を受賞した。

§2 研究実施体制

(1) 研究チームの体制について

(1)G0 グループ(研究機関別)
① 研究代表者:松本　裕治 (理化学研究所革新知能統合研究センター、チームリーダ)
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(研究機関別)
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(研究機関別)
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(研究機関別)
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(研究機関別)
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(研究機関別)
① 主たる共同研究者: 森　純一郎 (東京大学大学院情報理工学系研究科、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(研究機関別)
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

(8)G7グループ(研究機関別)
① 主たる共同研究者: 進藤 裕之(奈良先端科学技術大学院大学先端科学技術研究科、助教)
② 研究項目
・文書構造解析およびユーザインタフェース構築

(2)国内外の研究者や産業界等との連携によるネットワーク形成の状況について

G0 グループでは、CNRS の Pierre Zweigenbaum 博士、および、DFKI の Sebastian Möller 博士と共同研究を進めることで合意し、JST 戦略的創造研究推進事業 AIP ネットワークラボ「日独仏 AI 研究」の公募に「Knowledge-enhanced information extraction across languages for pharmacovigilance (医薬品安全性監視のための言語を越えた知識強化情報抽出)」というタイトルの共同研究提案を行い、採択された。医薬品の安全性に関する情報抽出に焦点を当て、エンティティ認識やエンティティ関係の抽出に関する本プロジェクトの成果を多言語展開することを計画している(研究期間: 2020 年 12 月～2024 年 3 月)。

G1 グループでは、フランスの HEC の Restrepo 准教授と法律データマイニングについての共同研究を行い、CREST の予算にて、 Legal Data Mining Conference (https://legaldatamining.com/)を 2019 年 3 月 21、22 日に行った。

G2 グループでは、井之上元助教を University of Southern California の知識表現の専門家 Andrew Gordon 氏のもとに約半年間、University College London (UCL) の Sebastian Riedel 博士のグループに約半年間派遣し、仮説推論機構に関する研究を実施した。その後も同グループの Pontus Stenetorp 博士と仮説推論機構についての研究を共同で行い、その成果は、言語処理学会年次大会言語資源賞受賞(85 件中 3 件)、トップ国際会議 ACL2020 に採択されるなど、顕著な成果を上げた。

G3 グループでは、国際インターンシップ学生の受け入れやサバティカルでの滞在等を通して、ドイツの University of Wuppertal の Bela Gipp 博士、Trinity College Dublin の Joeran Beel 博士、Technische Universitat Braunschweig の Wolf-Tilo Balke 博士、フランスの Nante 大学の Florian Boudin 博士、等と連携した。

G5 グループでは、大規模な引用ネットワークと文献のテキスト分析に基づく研究開発戦略立案支援を目的に、複数の国内民間製造企業と共同研究を実施した。具体的には、引用ネットワークの分析により技術シードとなる研究や重要な研究者を早期に特定することで、各企業の研究開発戦略の立案を支援した。また、大規模な文献情報データベースの利活用について国外の Elsevier 社や Springer Nature 社と意見交換を行うとともに成果の産業展開の検討を進めた。具体的には、文献情報データの共通フォーマット、論文のインパクト指標を含む新たなオルトメトリクス、などについて議論を進めた。