戦略的創造研究推進事業　AIP 加速研究
（AIP 加速 PRISM 研究）
研究課題「医療テキスト構造化のための
言語・知識処理基盤の構築」

研究終了報告書

研究期間　2018年8月～2021年3月

研究代表者：黒橋　禎夫
（京都大学大学院情報学研究科、教授）

§1 研究実施の概要

（1）実施概要

本研究では、医療分野における診療テキスト、患者テキストをターゲットとして、テキスト構造化のための言語・知識処理基盤の構築に関する研究開発を行った。本研究期間内においては特発性肺線維症（IPF）および肺がんに対する創薬に資するテキスト構造化を目標とし、その後においても本基盤がプレシジョン・メディシン等に資することを目標とした。

従来、我が国では病院や医療系研究機関を物理的に訪問しなければ患者の診療テキストを解析することができず、この状況では言語解析の研究を進めることが困難であった。そこで、本研究開始直後から医薬系研究者との研究会等を重ねて信頼関係を構築し、大阪大学医学部附属病院（国立研究開発法人 医薬基盤・健康・栄養研究所（以下、医薬基盤研）から提供）と国立がん研究センターの匿名化診療テキストをプロジェクト参加の大学研究室で利用できる環境を構築した。

その上で、医療テキストの性質を精査し、構造化の目標を明確化し、それに基づきコーパスへのアノテーション基準を策定し、肺線維症読影所見（約 600 件）、肺線維症診療録（約 300 件）、肺がん読影所見（約 2700 件）からなるアノテーション付与コーパスを構築した。構造化情報としては、医療エンティティに関して病名・症状、臓器・部位、検査、薬品、治療など、また医療エンティティごとに属性（病名・症状であれば有・無・疑など）を与え、さらに医療エンティティおよび時間表現間の関係性を与えた。当該コーパスは量・質ともに我が国の医療テキストコーパスとして卓越したものであり、国際的に見ても特定疾患に関するコーパスとして有数のものである。本プロジェクトの根幹となるこの部分については黒橋グループと荒牧グループの共同で実施した。

さらに、2018 年に発表された文脈言語モデル BERT をベースとした手法でこれらのコーパスを用いて医療テキスト構造化システムを学習し、肺線維症および肺がんの読影所見における医療エンティティの認識で 90%超、エンティティ間の関係認識で約 85%の精度を得た。

医薬基盤研では、肺線維症読影所見に対して本システムを利用してまず自動解析を行い、その結果を人手修正することで読影所見の構造化を行った。これに加えて診療録データ、血液検査、マルチオミックスデータを合わせて解析することで7つの標的候補分子を抽出し、医薬基盤研が有する統合データベース TargetMine などを使って精査することでさらに重要と考える蛋白質分子を推定することに成功している。今後優先順位をつけて検証実験が行われる予定である。

黒橋グループでは、さらに、医療テキストの時系列解析、実臨床データを用いた肺疾患の発症・予後に影響する薬剤の探索を行った。

荒牧グループでは、従来から当該グループで開発してきた大規模病名辞書「万病辞書」を、肺繊維症および肺がんに関して読影用語、用言句等 3000 語規模で拡張した。さらに、UMLS (Unified Medical Language System)等の国際標準オントロジーとの紐付けを行うことで PRISM 成果データの国際化に寄与した。また、上述のコーパス構築に加えて、一般公開されている症例報告や読影画像をベースとして元データを含めて公開可能な医療コーパスを整備した。

荒瀬グループでは、患者・家族からの視点である医療関連 SNS 等の患者テキストを対象として、テキストからの使用薬剤およびその奏功状況の自動認識技術を開発した。またその基盤となるデータセットとして、がんおよび IPF 患者の闘病ブログに対して薬剤奏功状況タグ付けコーパスを構築した。

戸次グループでは、医療テキストに対して理論言語学の最新の成果と深層学習を組み合わせた先端的な統語解析・意味解析・自動推論に関する研究を進めた。これにより、医療テキストの高度な構造化、柔軟な検索の実現可能性を示した。

これらの成果の多くは、言語資源に関する代表的国際会議 LREC、自然言語処理のトップ国際会議 ACL 等で発表を行った。また、本プロジェクトで開発した医療テキスト構造化のための言語資源（辞書、オントロジー、アノテーション付きコーパス、アノテーション基準マニュアル等）

および言語解析システムは、すべて研究利用可能な形で公開する予定である。

(2)顕著な成果

＜優れた基礎研究としての成果＞

1. 患者テキストにおける高精度かつロバストな情報抽出の実現

概要: 患者テキストの解析技術として、症状や投与薬剤およびその奏功をテキストから自動的に認識する技術を開発した。世界最高の認識精度を有すること、また低頻度の現象についてもロバストに認識できることを実験的に示した。また解析技術の基盤となるデータセットも構築している。肺がんおよび IPF 闘病患者ブログから抽出した 1,019 件の記事について、服用している薬剤とその奏功状況のタグ付けを行った。データセットの一部は研究利用に限定して Web にて公開している。この成果はトップ国際会議 ACL2020 で発表した。

2. 範疇文法と高階論理に基づく意味処理の診療テキストへの応用

概要: 診療テキスト論理推論システムのプロトタイプを開発した。本システムは CCG 構文解析プログラム、高階論理を用いた意味合成プログラム、および自動推論プログラムを統合したものであり、時系列情報を含む診療テキストへの自然言語での問い合わせに対して適切な検索結果を返すものである。本研究のうち CCG 構文解析の診療テキストドメイン適応に関わる研究論文は、トップ国際会議 ACL2019 に採択された。

3. 動的イベント表現とマルチタスク学習による時間関係抽出技術の向上

概要:文書中の事象の時間関係の特定は、事象の前後関係や継続期間・頻度などを理解する上で重要な技術だが、その精度はまだ十分ではない。本研究では動的イベント表現とマルチタスク学習を利用した新しいニューラルモデルを提案し、時間関係抽出の性能を向上させた。英語コーパスと日本語コーパスの両方において、提案モデルが最先端のモデルを大差で凌駕することを包括的な比較実験で示した。この成果は当該分野のトップ会議 EMNLP 2020: Findings volume に採択された。

＜科学技術イノベーションに大きく寄与する成果＞

1. 医療テキストコーパスの構築

概要: 医療文書に対する自然言語処理(医療言語処理)の基盤となる、アノテーション付コーパスを構築する研究開発を行った。肺疾患を扱う診療録と読影所見に対し、病変や部位といった医療表現エンティティと、エンティティ同士の関係を定義し、3500 件以上の医療文書にアノテーションを付与した。このデータの一部および策定されたアノテーション仕様は公開を予定しており、国内の医療言語処理研究促進に寄与する。

2. 医療情報抽出システムの構築

概要:構築した医療テキストコーパスを用いて、医療エンティティ・属性・関係認識システムを構築した。深層学習言語モデル BERT を用いた Joint モデルを採用し、医療エンティティ(症状、部位など)および文内・文間のエンティティ間関係の認識 F1 精度が、肺がん読影所見でそれぞれ 97.12、93.00、肺線維症診療録で 86.86、69.94 と、ベースラインの pipeline モデルを大きく上回る精度を達成した。

＜代表的な論文＞

1. Shuntaro Yada, Ayami Joh, Ribeka Tanaka, Fei Cheng, Eiji Aramaki, Sadao Kurohashi. ″Towards a Versatile Medical-Annotation Guideline Feasible Without Heavy Medical Knowledge: Starting From Critical Lung Diseases,″ In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), pp. 4565-4572, Online, (2020).

概要: 読影所見や診療録など種々の医療文書に対応した汎用の医療表現アノテーション仕

様を提示した。コーパスの記述様態を安定させるため、死因の多くを占める肺疾患に着目した上で、医学知識に乏しい作業者によるアノテーションを可能とするための工夫を導入して仕様の頑健化とコスト低減を実現した。提案アノテーションを施した約 1100 件の医療文書での固有表現抽出タスクで F1 値 95.3 を記録し、提案仕様が大規模臨床 NLP プロジェクトにも適用可能であることを示した。

2. Sora Ohashi, Junya Takayama, Tomoyuki Kajiwara, Chenhui Chu, Yuki Arase. ″Text Classification with Negative Supervision,″ In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL2020), pp. 351-357, Online, (2020).

概要: ソーシャルネットワーキングサービスに投稿される文から、投稿者が罹患している疾病を自動的に認識する手法を開発した。異なるラベルを持つ文が異なるベクトル表現を持つようにマルチタスク学習を導入することで、表層が類似した文の誤分類を抑制する。文分類だけでなく文書分類、またラベル数に関わらず有効であり、日本語・英語・中国語の 3 言語において効果的な手法であることを示した。

3. Masashi Yoshikawa, Hiroshi Noji, Koji Mineshima, Daisuke Bekki. ″Automatic Generation of High Quality CCGbanks for Parser Domain Adaptation,″ In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL2019), pp. 129–139, Florence, Italy, (2019).

概要:CCG 統語解析器のドメイン適応技術を開発した。CCG bank の文の係り受け木を入力、CCG 木を出力とするデータで訓練したニューラルネットを用いて、対象となるドメインのテキストの係り受け木から CCG 木を自動生成し、統語解析器を再訓練する。生物医学ドメインを含む複数のドメインで大幅な解析精度向上を確認した。

§2 研究実施体制

(1)研究チームの体制について

①黒橋グループ
研究代表者:黒橋　禎夫 （京都大学大学院情報学研究科　教授）
研究項目:
・疾患知識ベースの構築
・医療テキストの知識処理
・実臨床データを用いた肺疾患の発症および予後に影響する薬剤の探索

②荒牧グループ
主たる共同研究者:荒牧　英治 （奈良先端科学技術大学院大学研究推進機構　教授）
研究項目:
・コーパスの構築
・オントロジーの構築
・自動構造化システムの構築

③荒瀬グループ
主たる共同研究者:荒瀬　由紀 （大阪大学大学院情報科学研究科　准教授）
研究項目:
・臨床テキストの解析
・患者テキストの解析

④戸次グループ
主たる共同研究者:戸次　大介 （お茶の水女子大学基幹研究院　准教授）
研究項目:
・医療テキストの構文解析
・医療テキストの自動推論

(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

本プロジェクトは官民研究開発投資拡大プログラム（PRISM）「新薬創出を加速する人工知能の開発」の一貫として、医薬基盤研、国立がん研究センター、産業技術総合研究所等との協力で進められた。このこともあり、本プロジェクトの研究会には医薬基盤研、国立がん研究センターの研究者にも常時参加頂いた。

また、黒橋と荒牧が国立がん研究センターでの招待講演を行い、その結果として両名が国立がん研究センター客員研究員となるなどして協力関係を深めた。

2019 年 9 月には情報系の国内最大規模の会議 FIT2019 において、本プロジェクトおよび PRISM 参画研究者が中心となってイベント企画セッション「医療と自然言語処理のこれから」を開催し、情報系研究者と医薬系研究者の相互理解の促進に務めた。

2021 年度からは厚生労働科学研究費の枠組において引き続き協力体制を継続している。