2018 年度
AIP-PRISM 実績報告書

黒橋　禎夫

京都大学大学院情報学研究科
教授

医療テキスト構造化のための言語・知識処理基盤の構築

§1. 研究成果の概要

本研究では、ビッグデータ基盤領域 CREST「知識に基づく構造的言語処理の確立と知識インフラの構築」の成果を発展させ、医療分野における臨床テキスト、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・知識処理基盤を構築することを目的とする。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する。プロジェクトの第一年度となる今年度は、各研究項目について以下の成果を得た。

医療用語オントロジーと医療表現アノテーション・コーパスの構築：荒牧グループ（奈良先端科学技術大学院大学）

医療テキスト解析の基盤となる医療用語について整備し、一部オントロジー化を行った。まず、病名・症状名とその標準化コード（ICD コード）からなる大規模病名辞書「万病辞書」を拡張し、本研究の対処疾患である肺線維症に関する病名・症状名（以下、単なる病名）について細分化した情報を加えた。現在、362,866 件の病名用語（うち、25,678 件が標準病名）が収載され、標準病名または人手でのコーディングが行われた病名は 73,342 件となっている。同時に、関連する医薬品についても薬品と成分のリストの整備を行っており、薬剤添付文書を中心に数万規模の表現を整理している。肺部位を中心に人体部位についても 400 表現を集め、部分全体関係を記述している。

また、コーパスについても、国立がんセンターに提供いただいた読影所見を中心に、国立がんセンター、京都大学と頻回に議論しながらガイドラインを作成し、1000 件の文書について表現のアノテーションを行った。これは機械学習の教師データとして用いることができる規模である。

医療テキストからの情報抽出の高度化：鬼塚グループ（大阪大学）

表現の多様性に頑健な病名・症状の認識手法の実現に向けて、テキストのベクトル表現生成技

術を応用した患者テキスト解析手法の開発に着手した。対象とする患者テキストに近い特性をもつNTCIR データを用い、テキストから記述者の対象疾患・症状の有無を予測する問題に取り組んだ。テキストのベクトル化は現在活発に研究が行われている分野で数多くの手法が提案されているが、医療ドメインにおける有効性は明らかとなっていない。そこで今年度は (1) 既存のテキストベクトル化技術の性能評価と分析、および (2) 患者テキスト解析の基礎的なモデルの開発に取り組んだ。提案モデルは NTCIR データにおいて最高性能を達成している。

医療テキストのための表現計算モデルの構築:戸次グループ(お茶の水女子大学)

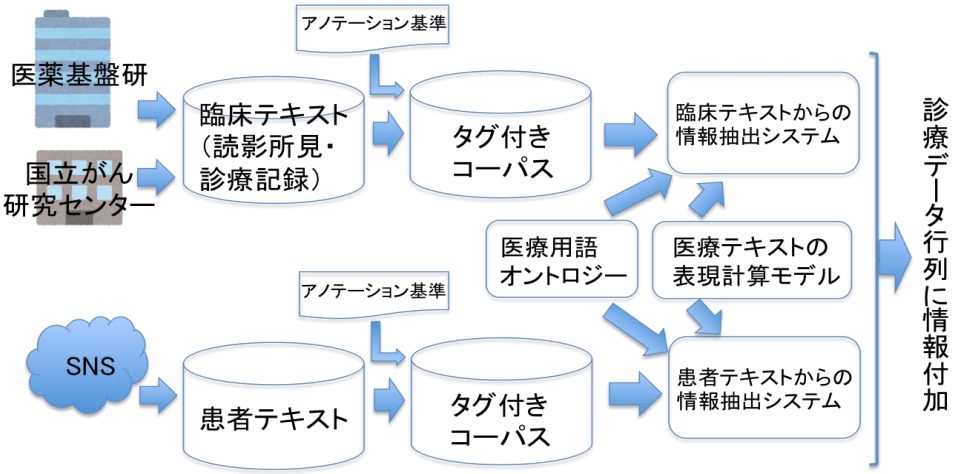
医療テキストに対して先端的な意味解析を行い、医療テキストの整理・構造化に応用することを目指して、【医療テキストの構文解析】と【医療テキストの自動推論】という二つの小項目について研究を進めた。前者においては、組み合わせ範疇文法(Combinatory Categorial Grammar: CCG)に基づく頑健かつ高精度な構文解析器 depccg を、医療テキストツリーバンクの構築を通して、医療テキストドメインに適応させる手法を開発した。後者においては、depccg の出力を高階論理表現に変換して自動推論を行う意味処理システム ccg2lambda を用いて、医療テキストに対するイベント時系列解析、類似症例検索、事実性判定といった高度な意味処理を実現することを目指しており、本年度は時制・時間関係情報の付与に取り組んだ。

疾患知識ベースの構築と医療テキストの知識処理:黒橋グループ(京都大学)

非文法的、断片的に専門用語が羅列される医療テキストから、症状名、その事実性、判断の根拠などを構造的に抽出するための基盤として、知識ベースを構築することを目指している。

本年度は、情報抽出に着手するにあたりコーパスが必要となることから、荒牧グループと共同で電子カルテデータを分析しコーパスのタグ付与基準を策定した。構築されたコーパスを利用し、深層学習モデル BERT を用いて医療表現の自動認識を行い、高い精度を達成した。さらに、電子カルテを時系列に分析するために時間表現解析が重要となることから、ニューラルネットワークに基づく日本語の時間表現解析の研究を進めた。

また、米国有害事象セルフレポート 650 万症例に出現する 55 万種類の医薬品名を 5 千の有効成分名に名寄せし、IPF 発症率に影響する交絡因子を抽出した。



研究成果の概要図

§2. 研究実施体制

(1)黒橋グループ

① 研究代表者:黒橋　禎夫 （京都大学大学院情報学研究科　教授）

② 研究項目

・疾患知識ベースの構築

(2)荒牧グループ

① 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情報科学研究科　特任准教授）

② 研究項目

・オントロジーの構築

・コーパスの構築

(3)鬼塚グループ

① 主たる共同研究者:鬼塚　真 （大阪大学大学院情報科学研究科　教授）

② 研究項目

・患者テキストの解析

(4)戸次グループ

① 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）

② 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