2019 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、ビックデータ基盤領域 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。プロジェクトの第二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）

本年度の主な成果はコーパス構築である。昨年度に引き続き，国立がんセンターに提供いただいた読影所見および阪大病院から提供いただいた診療録（総計 1200 件程度）に対し、医療表現のアノテーションを実施した。昨年度が病名や部位名、その修飾語といった基本的な医療表現を対象としていたところ、本年度は医薬品や入退院状況、時間表現などを含めた広範かつ統合的なアノテーション仕様として更新してのものである。この成果は国際会議 Language Resources and Evaluation Conference 2020 で発表予定である（京都大学グループとの共同）。さらに本年度は、より精細な情報抽出のために、医療表現間の関係を付与する仕様も策定し、同じコーパスにアノテーションを実施した。以上の成果は国立がんセンターおよび京都大学との頻回にわたる議論に基づく。

アノテーションされたコーパスから病変等の情報を抽出し、表形式に構造化するシステムも開発した。京都大学グループの機械学習モデルと組み合わせることで、アノテーションされていない読影所見からも情報抽出できる成果となる。

医療用語オントロジーについては、国際化に向けた予備的な検討を実施した。国際標準の医学概念シソーラス（ICD10 や TNM 分類（腫瘍分類））にマッピングすることで、オントロジーの可用性向上を見込めることを確認した。

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）

高度な患者テキスト解析手法の実現に向け、本年度は（1）闘病ブログに対する薬剤奏功状況アノテーションおよび（2）高精度な疾病認識モデルの開発、に取り組んだ。(1) では、肺がん患者の公開闘病ブログを Web から収集し、投与された薬剤名、薬剤の奏功状況、ICD10 コードおよび MedDRA コードの付与、奏功状況カテゴリの付与、を行うアノテーションを実施した。合計で 684 件の記事に対するアノテーションを完了しており、一部アノテーションデータは Web サイト[1]にて公開している。（2）では、意味的に類似しているが分類ラベルが異なるテキストにおいて分類性能が低下する現象に着目し、深層学習によるテキスト分類モデルに対しマルチタスク学習を導入した。異なるラベルを持つテキストのベクトル表現間の類似度が下がるよう負の教師信号を与えることで、テキストベクトル化モデルを補正し、ロバストな分類を可能とする。これにより、SNS テキストからの疾病認識を含む 9 種類の分類タスクの内、8 つのタスクで世界最高性能を達成した。

これらの成果は Annual Conference of the Association for Computational Linguistics (ACL2020) および International Conference on Language Resources and Evaluation (LREC2020) で発表予定である。

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、【1.医療テキストの構文解析】【2.医療テキストの自動推論】【3.医療テキストの談話関係】という三つの小項目について研究を進めた。1.においては、CCG 構文解析器 depccg を医療テキストドメインに適応させる手法をトップカンファレンスにおいて発表した。2.においては、医療テキストに対するイベント時系列解析の実現を目指し、時間関係認識のための基礎技術を構築した。3.については、日本語の談話関係のための理論体系を新たにデザインし、アノテーションスキーマを確立した。2.と 3.の研究は、来年度以降の研究において、類似症例検索、事実性判定といった高度な意味処理の実現につなげる予定である。

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）

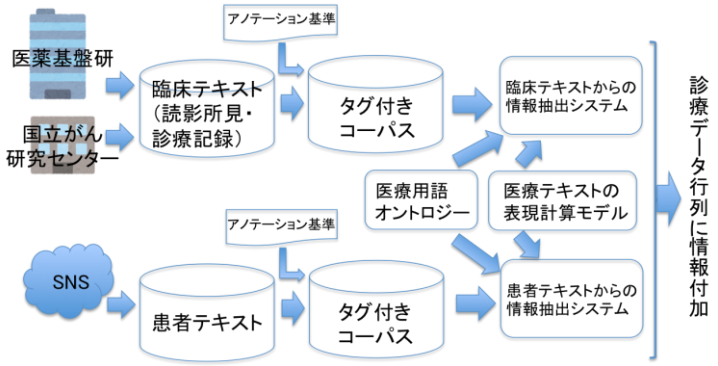
非文法的、断片的に専門用語が羅列される医療テキストの構造化に向けて、(1)疾患知識ベースの構築、(2)知識に基づく医療テキスト解析に取り組んだ。

(1)については、昨年度に引き続き情報抽出に必要なコーパス構築に荒牧グループと共同で取り組んだ。対象とする医療表現の範囲を拡張し、さらに医療表現間の関係アノテーションの仕様も策定した。また、得られた医療表現アノテーションデータと、本グループで昨年度構築した高精度の医療表現自動認識システムを用い、半自動アノテーションによるコーパス構築も進めた。これらの

[1] https://yukiar.github.io/adr-jp/

成果は荒牧グループと共同で LREC2020 で論文発表を行う。

(2)では、臨床テキストの時系列情報の解析に向けて時間的順序関係推定の精度向上に取り組み、英語の Timebank-dense と日本語の BCCWJ-Timebank の両データセットにおいて既存手法を大きく上回る精度を実現した。この成果は国際会議に論文投稿中である。さらに、構築したコーパスを利用して医療表現認識・関係推定システムを構築し、高い精度を達成した。



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

・医療テキストの知識処理

(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