別添４

厚生労働科学研究費補助金（政策科学総合研究事業）
（臨床研究等 ICT 基盤構築・人工知能実装研究事業）
分担研究報告書

課題名　　：新薬創出を加速する症例データベースの構築･拡充/創薬ターゲット推定アルゴリズムの開発に関する研究

研究分担者名：荒牧 英治
国立大学法人奈良先端科学技術大学院大学 先端科学技術研究科 教授

研究要旨
本研究では、「創薬ターゲットの枯渇問題」を克服すべく、電子カルテを始めとする診療テキストから患者の情報を抽出する技術を研究開発するにあたり、症例報告や読影所見から重要な表現を自動抽出し、国際的なコードに変換する自動構造化技術を開発する。この基盤であり中核をなすオントロジーと、このオントロジーを評価するためのテストベッドの整備に従事してきた。すでにオントロジーは完成し、公開準備を進めている。またテストベッドとして、データセットと可視化システムを開発した。

A．研究目的

本研究では、「創薬ターゲットの枯渇問題」を克服すべく、電子カルテを始めとする診療テキストから患者の情報を抽出する技術を研究開発する。このために、症例報告や読影所見から重要な表現を自動抽出し、国際的なコードに変換する自動構造化技術を開発する。これは次の課題の解決が必要となる。

◎オントロジーの整備

国際的なコードには用語の不足も多いことから、コードの拡充、整備を行う。これまで読影所見に頻出する 1000 用語の整備を行った.

◎オントロジー評価テストベッドの整備

オントロジー単体ではその評価は困難で、使用目的（実応用）を決めて初めて実際的な評価が可能となる。そこで、類似症例検索システムのプロトタイプや病名抽出タスク向けデータを構築し、本オントロジーを用いた手法の精度検証に用いられる環境（テストベッド）を準備する。

B．研究方法

【オントロジー】

本グループは、総括班の研究推進に必須となる医学用語リソース（本研究ではオントロジーと呼んでいる）とコーパス（機械可読な情報が付与されたテキストデータ）の２つのリソースを構築してきた。オントロジーは、後述するコーパスから収集された用語に標準医学表現を紐付けたものである。コーパスについては黒橋グループ、荒瀬グループと相談の上、3000 件以上の医療文書に対し、医学的な固有表現などのそれら同士の医学的関係の付与（アノテーション）を行ってきており、拡充を続けている。コーパスからのオントロジー作成にあたり、コーパスに出現する病変や症状、部位を表す表現がアノテーションの結果よりわかるので、リストアップされた表現を目視で精査しながら国際コードを付与した。本オントロジーに収録する表現は、目標として設定した類似症例検索に役立ちうる範囲とした。

【テストベッド】

本オントロジーを活用することで性能等が向上するシステムとして、症例中の患者病態推移を時系列で直感的に可視化を試みた。開発したシステムは、近年の、クリニカルパスに基づくチーム医療の重視に資するシステムと言え、臨床現場のコミュニケーションを促進できる。可視化システムのコンポーネントには、病名抽出器や時間表現抽出器が含まれ、これら要素技術にも本オントロジーを用いることができる。加えて、オントロジーとも共通するアノテーション仕様を適用し、臨床テキストからの患者情報抽出の性能を評価できる汎用でオープンなコーパスデータ（「評価用コーパス」）を構築する。

（倫理面への配慮）
本研究は、医薬基盤・健康・栄養研究所において倫理審査、承認を得た後、人を対象とする医学系研究に関する倫理指針に従って実施している。上記研究は、個人情報が削除済みのテキストデータ、及び荒牧グループが作成・保有している模擬コーパスにておこない、個人情報保護の観点からは安全なデータである。

C．研究結果
【オントロジー】
肺がん・肺線維症に関する用語を網羅的に収録したオントロジーである「PRISM Lung Disease Ontology」が完成した。計算機で処理しやすいCSV及びJSON形式のデータとした。オントロジーの元となったコーパスの構築時に作業者が取り組むアノテーション作業を説明したガイドラインを日本語で作成していた。これを英語に翻訳し、日英とも DOI 付のデジタル情報資源として公開することで、本研究のアノテーション仕様に基づくコーパス作成を関連研究者や実践者も実施できるような環境を整えた。さらに，アノテーション仕様の策定や実施に関する詳細も含め、作成したコーパスについて報告した論文をオープンアクセスで雑誌『自然言語処理』に投稿したことで、同様の医療言語処理に取り組む研究者へのノウハウ共有に貢献した。

【テストベッド】
症例の時系列可視化システムプロトタイプ（ベースライン）　「HeaRT」が開発済みであり、現在、国内で特許として申請し、審査を受けている。可視化による臨床業務支援効果を検証するため、2023 年度にはNEC 社の協力で病院での実証実験を進める運びとなった。
「評価用コーパス」として、J-STAGE でオープンアクセスのもと公開されている症例報告論文 224 件から作成済みのものと、同じ読影画像に対して複数の読影医が執筆した読影所見 135 件のものがある。病名等のアノテーションは本コーパスと同様の仕様に基づく。これを英語に翻訳してテストベッドとしての一般性を高めた。症例報告は中国語にも翻訳したほか、読影所見はさらに 224 件を追加した。

D．考察
【オントロジー】
プライバシーポリシーの調整も終え、公開可能な状態。現在、公開場所、タイミングを選定中である。

【テストベッド】
類似症例検索及び症例時系列可視化のプロトタイプ（ベースライン）を開発できた。後者は特に，プロジェクト成果物であるアルゴリズムをクラウド上で公開する「峰」プラットフォーム（医薬基盤・健康・栄養研究所の提供）からデモとして公開されるため、多くの関連研究者・開発者の目に留まることが期待される。これらをベースに、本オントロジーを活用するコンポーネントを追加した新規システムを開発すれば、プロトタイプとの性能比較によって本オントロジーの間接的（かつ実際的）評価も可能になる。
さらに、病名標準化タスク向けの新規「評価用コーパス」も、症例報告と読影所見の 2 つを作成できた。このコーパスを用いて、医療言語処理のシェアードタスクである Real-MedNLP を、国際ワークショップである NTCIR-16 の傘下で企画・運営した。本プロジェクトのアウトリーチも兼ねており、最終的な参加は10 チームと、NTCIR-16 で採択されたタスクの中でも大きな注目を集めた。

E．結論
症例報告や読影所見から重要な表現を自動抽出し、国際的なコードに変換する自動構造化技術を開発するために必須となる、「オントロジー整備」と「テストベッド開発」を進めた。オントロジーは完成し、公開を控えた準備段階にある。また、テストベッドとして 2 つのシステムと 2 つのデータセットを提案・開発できた。プロジェクト終了後も、オントロジーの正式な公開とテストベッドの拡充を進める。

F．健康危険情報
該当せず

G．研究発表

１．論文発表

1) 矢田 竣太郎, 田中 リベカ, Fei Cheng, 荒牧 英治, 黒橋 禎夫：汎用的な臨床医学テキストアノテーション仕様およびガイドラインの策定：重篤肺疾患ドメインに着目して, 自然言語処理, 29(4), pp. 1165-1197, 2022 (2022/12/15)

２．学会発表

1) Lean Franzl Lim Yao, Kongmeng Liew, Shoko Wakamiya, Eiji Aramaki: Extracting Spatio-Temporal Trends in Medical Research Prioritization Through Natural Language Processing of Case Report Abstracts, MedInfo 2023 (2023/7/10-12, Sydney, Australia)
2) Faith Wavinya Mutinda, Kongmeng Liew, Shuntaro Yada, Shoko Wakamiya, Eiji Aramaki: PICO Corpus: A Publicly Available Corpus to Support Automatic Data Extraction from Biomedical Literature, In Proceedings of the Workshop on Information Extraction from Scientific Publications (WIESP 2022), 2022 (2022/11/21, Online)
3) Fei Cheng, Shuntaro Yada, Ribeka Tanaka, Eiji Aramaki, Sadao Kurohashi: JaMIE: A Pipeline Japanese Medical Information Extraction System with Novel Relation Annotation, In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022) (Poster), pp. 3724-3731, 2022 (2022/6/22, Marseille, France)

H．知的財産権の出願・登録状況（予定を含む。）

１．特許取得

なし

２．実用新案登録

なし

３．その他

なし