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、2 つのシステムと 2 つのデータセットを開発できた。今後はテストベッドを用いたオントロジー評価実験を中心に進める予定である。

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が、対象を限定すると、不足や不要な部分も多い。一方で何ら制約や目標を設定せずに整備を進めるのも困難であることから、実用的なアウトカムたりうる類似症例検索をゴールとして整備を進める。

◎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課題やシステムを設計する。

まず、オントロジー構築にあたり目標とした類似症例検索システムについて、古典的な手法を用いてプロトタイプ（ベースライン）を構築した。 これはオントロジーを用いないものとし、今後開発予定のオントロジー活用手法がこれより優れることをもってオントロジー自身の間接的な評価に用いることのできる「テストベッド」である。

類似症例検索のような医学文書処理の一環として、症例中の患者病態推移を時系列で直感的に可視化することを考えた。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、実現すれば、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。そこで、もう一つの「テストベッド」として、症例文書の時系列可視化システムを開発した。これも類似症例検索システムの場合と同様に、オントロジーを使わない手法に基づくプロトタイプ（ベースライン）とする。

オントロジーの活用が期待される実応用の 1 つに病名正規化（標準化）がある。このタスクは、テキストに出現する病変や症状の自然言語文記述を同定した上で、その記述が指す医学概念を国際コードに紐つけるというもので、医師が書くテキストにおいて同じ病変を表しながらも表記が異なる記述から一貫した情報抽出を可能とするために必須の技術である。このタスクで、本オントロジーを用いた手法が、他の手法より優位であることなどを今後示すことができれば、「テストベッド」とすることができる。ただし、本研究で構築したコーパスはこのタスクに利用できるが、本オントロジーもまたこのコーパスから構築されており、オントロジーの評価に向かない。そこで、本コーパスとは異なる文書セットに本コーパスと同様のアノテーションを施した、新規の「評価用コーパス」を構築する。再現性の観点から、評価用コーパスは一般公開可能なライセンスの文書を対象とした。

（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。 1197 エントリからなる。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。

【テストベッド】

プロトタイプ（ベースライン）の類似症例検索システムを実装した。これは症例文書を単語出現頻度に基づくベクトル表現に変換し、文書ベクトル間の cos 類似度の高さでもって類似症例とみなすものである。ただし、考慮する単語は、黒橋グループ開発の固有表現抽出器による病変や部位等の医学的クラスの識別情報でフィルタリングできるようにした。https://aoi.naist.jp/prism-search/ からアクセスできる。

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。また、海外特許としての申請を目指し、JST からの支援を受けるための手続きも開始した。

「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。一部を中国語にも翻訳している。

D．考察

【オントロジー】

オントロジー自体は完成したが、公開にあたり、プライバシーポリシーの調整が必要である。現在、開発関係者を交えて議論を進めている。

【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。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。ただし、類似症例検索については一般の検索システムと同様の評価手法を用いることができるが、症例時系列可視化については標準的な評価手法が設定されていないため、その有効な評価手法を発案・設計することが求められる。

また、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・運営している。本コーパスや本オントロジー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。6 月の NTCIR-16 会議に向けた準備を進めている。

「テストベッド」はこれまでに開発したもの以外にも考えられる。病変に関するテキスト上の記述は、患者が専門用語を使わずに書くような動詞や修飾語からなる表現に顕著であるが、「何文字目から何文字目までが病名である」との判断に難しい場合がみられる。一方、病名抽出にこれまで用いてきた固有表現抽出（NER）はそういった厳密な境界を決めることが前提となっている。そこで、臨床医学表現について境界の曖昧性を許容する新規の固有表現抽出タスク「Fuzzy NER」について検討を進めている。各文について必ず 1 つの固有表現を含む単位に分割することで、後段に病名正規化タスクを置けば、文中の大まかな単位で病変・症状の出現箇所がわかるのではないかと考えられる。すなわち、これも「テストベッド」の一つとできる見込みがあることから、タスク設計を継続する。

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。今後はオントロジーの正式な公開とアウトリーチ、新規テストベッドの開発や、開発済みテストベッドを用いたオントロジーの評価を実施する。

F. 健康危険情報

該当せず

G．研究発表

1．論文発表

1) Faith Wavinya Mutinda, Shuntaro Yada, Shoko Wakamiya, Eiji Aramaki: Semantic Textual Similarity in Japanese Clinical Domain Texts Using BERT, Methods of Information in Medicine, S 01, pp. e56-e64, 2021. (https://doi.org/10.1055/s-0041-1731390)
2) 荒牧英治：自然言語処理の医療への応用，先進医療 NAVIGATOR 医療と AI 最前線 新刊，2022 (2022/1/31)

2．学会発表

1) Faith Wavinya Mutinda, Shuntaro Yada, Shoko Wakamiya, Eiji Aramaki: AUTOMETA: Automatic Meta-Analysis System Employing Natural Language Processing, MedInfo 2021 2021 (2021/10/2-4).
2) Yuta Nakamura, Shouhei Hanaoka, Yukihiro Nomura, Naoto Hayashi, Osamu Abe, Shuntaro Yada, Shoko Wakamiya, Eiji Aramaki: Clinical Comparable Corpus Describing the Same Subjects with Different Expressions, MedInfo 2021 2021 (2021/10/2-4).
3) 氏家翔吾，磯颯，荒牧英治: 文脈化埋め込み表現を用いた対照学習による病名正規化，言語処理学会第 27 回年次大会（NLP2021）（オンライン），2021 (2021/3/16).

H．知的財産権の出願・登録状況（予定を含む。）

1．特許取得

1) 矢田竣太郎，荒牧英治：臨床テキスト情報時系列データ作製方法及び装置、並びに、臨床テキスト情報時系列可視化表示方法及び装置、並びに、臨床テキスト情報時系列可視化システム （特願 2021-165067）(2021/10/6)

2．実用新案登録

なし

3．その他

なし