

科学研究費助成事業 研究成果報告書

平成 26 年 5 月 2 日現在

機関番号：82602

研究種目：若手研究(A)

研究期間：2012～2013

課題番号：24680027

研究課題名(和文) 疾患知識ベースの効率的な整備手法と利用促進に関する研究

研究課題名(英文) Research on efficient compilation and promotion of disease knowledge base

研究代表者

奥村 貴史 (Okumura, Takashi)

国立保健医療科学院・その他部局等・特命上席主任研究官

研究者番号：50553400

交付決定額(研究期間全体)：(直接経費) 14,000,000円、(間接経費) 1,140,000円

研究成果の概要(和文)：高品質な医療用情報技術の研究開発には、疾患の取りうる所見や検査結果などを収載した疾患知識ベースが求められる。そこで本研究は、重要性の高い疾患について手動編纂したデータと、自動生成した希少疾患データを組み合わせることで、実用的な精度を有する疾患知識ベースを効率的に整備する手法を検討した。また、こうして構築した疾患知識ベースを品質管理するために、利用者からのフィードバック機構を備えた診断支援システムを開発した。これらの研究の結果、網羅性の高い疾患知識ベースと実用的な水準の診断支援システムの整備に加え、各種の医学文献における疾患知識の効率的な抽出に関する知見が得られた。

研究成果の概要(英文)：Improvement of information systems for clinical practices requires a disease knowledge base that contains possible clinical findings and laboratory examination results for each disease. However, compilation of such knowledge base necessitates considerable time and money. Accordingly, this study proposed to hybridize the data for major diseases, manually compiled to guarantee necessary quality, and the data for rare diseases, automatically generated for broader coverage, in the aim of efficient compilation of practical disease knowledge base. Additionally, toward quality control of the resulting knowledge base, a diagnostic decision support system is built with a feedback mechanism to collect usage information to efficiently improve the quality of the disease knowledge base. The research yielded disease knowledge base and a diagnostic decision support system with practical quality, as well as various knowledge about automatic acquisition of disease knowledge out of medical literature.

研究分野：総合領域

科研費の分科・細目：図書館情報学・人文社会情報学

キーワード：医療情報 疾患知識ベース 診断支援システム

## 1. 研究開始当初の背景

医療用情報システムは、医療現場の適切な情報化を通じて、医療の効率化と医療水準の向上に寄与することが期待されている。しかしながら、現在の医療用情報システムは、極めて品質が低いものに留まっており、医療従事者の勤務負担を押し上げている。こうした状況を改善し医療用情報システムの品質向上を果たすためにはさまざまな施策が求められるが、その一つとして診断支援技術の向上がある。

診断支援技術は、臨床において時折生じる診断困難症例の診療に際して、患者と医療従事者双方のさまざまな負担を軽減することができる。また、誤診等の防止へと応用することにより、医療安全の向上をもたらす。これは、間接的に医療現場の負担を軽減する。さらに、診断支援技術は、患者状態の自動的な把握を可能とし、これは、指示入力の支援などを通じてユーザーインターフェースの大幅な効率化をもたらす。

こうした診断支援技術のメリットを享受するためには、まず、臨床的な診断を行うための知識処理技術が求められる。しかしながら、それ以上に重要な要素として、あらゆる疾患に関する症状や検査結果についての情報が機械可読な形で整理されている必要がある点が挙げられる。この情報がなければ、対象患者の状態を自動認識することができない。診断支援技術の研究開発においては、この疾患に関する情報の集合、すなわち、疾患知識ベースが、欠かすことのできない要素となる。

しかしながら、疾患知識ベースの編纂には医師の関与が不可欠であり、一般的な知識ベースと比してデータ生成のコストがとりわけ高いという問題がある。また、あらゆる疾患を網羅するうえでは、遺伝疾患から感染症まであらゆる既知疾患についての情報整理を行う必要がある。これらの結果、疾患知識ベースの開発には、膨大な時間とコストを要するため、診断支援技術の研究開発の障害となっていた。

## 2. 研究の目的

本研究は、このような背景を受け、疾患知識ベースを効率的に整備する手法の研究開発を目的とする。その最大の課題は、いかにして数多くの疾患に関する情報を効率的に収集するかという点にある。そのためには、近年発達している自然言語処理を用いることで、さまざまな医学文献から自動的に疾患情報を収集する方向性が考えられる。しかし、自然言語処理による自動処理には、情報を誤って解釈したり、重要な情報を拾い損ねたりといった可能性があり、そのまま臨床応用するには問題が多い。

この問題を回避するうえで、データの一部をサンプル調査にすることで精度管理のコストを軽減する方法がありうる。しかしながら、診断支援というシステムの目的上、サンプルの漏れにより生じる損害をコスト上の理由で正当化することは、社会的に困難であろう。そこで、より現実的な手法として、疾患の医学的重要性や出現頻度を加味し医学的な意義の大きい疾患を重点的に監査することで、実用性を保ちつつコストの軽減を実現する戦略が求められることになる。

本研究では、この重要疾患に対する精度重視の疾患知識ベースと、網羅性を重視した自動生成による疾患知識ベースのハイブリッド化を提案する。すなわち、誤りが許されない急性疾患や頻度の高い疾患等に対しては全数を手動編纂し、網羅性が求められる希少疾患については自動生成を行う。さらに、自動生成した疾患情報について精度管理を行うために、生成した疾患知識ベースを利用する診断支援システム側に診断結果に対するフィードバック機構を設けることで、疾患知識ベースの精度を自律的に向上させていく手法を提案した。

このように、コンテンツの品質管理をコミュニティに委託するモデルは、**Community-driven curation** と称される。たとえば、**Wikipedia** は、利用者が勝手に記事を追加し編集することが可能であるが、利用者相互の監査によって、大規模な百科辞典をそれなりの品質で構築し、実用性を備えることに成功している。しかし、**Community-driven curation** は必ずしも成功例だけではない。とりわけ、本研究は情報整備のコストを下げることで大きな目的であるために、データ提供者に金銭的なインセンティブなどを設けることが出来ず、より難度が高いと言える。一方で、**Community-driven curation** を用いて低コストに疾患知識ベースを充実させることができれば、各種の医療用情報技術を実用化させていく基盤が実現する点でも研究的価値が高いと考えられる。

## 3. 研究の方法

疾患に関する知識ベースの中核は、疾患と所見の関係性情報にある。この情報の編纂に際しては、**ICD** 等の標準的な疾患分類を疾患マスタとしたうえで、それぞれの症状を **MeSH** などの標準を用いて記述するという方法が考えられる。しかしながら、**ICD** は疾患の分類体系であるために、臨床上区別が付かない疾患が別疾患として記載されていたり、同一疾患が複数のコードに分かれる等、診断支援システムが依拠する疾患単位として必ずしも理想的なものではない。また、診断支援システムは、理学所見に加えてさまざまな臨床検査結果やリスクファクタ等を扱わなければならない、標準語彙だけでは十分な表現力を備えているとは言えない。

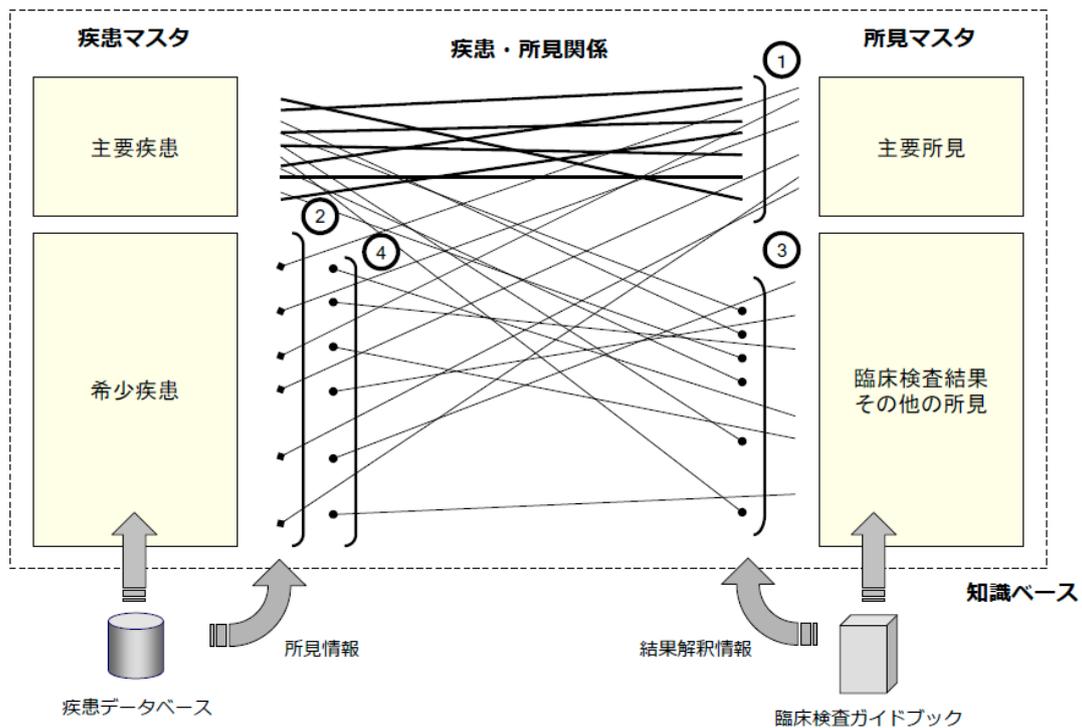


図1 疾患知識ベース構成図

そこで、本知識ベースでは、独自の疾患マスタと所見マスタを定義したうえで(図1-①)、自動生成のために希少疾患と臨床検査結果所見のエントリをそれぞれ追加し、疾患・症状関係の自動生成を試みた(図1-②~③)。自動生成に際しては、著名な疾患データベースである OMIM (Online Mendelian Inheritance in Man)、Orphanet に記載されている疾患、それぞれ 6000 強を対象とし、各疾患の説明文より所見情報、並びに、臨床検査結果情報の抽出を試みた(図1-②、④)。さらに、臨床検査のガイドブックに記載されている臨床検査それぞれについて、検査結果と原因疾患のペアに関する情報抽出を行い、知識表現を試みた(図1-③)。また、診断に際しての参考とするために、疾患マスタ上の主要疾患それぞれに対して、ウェブ上の参考文献情報を網羅的に収集すると共に、その参考文献データを教師データとした疾患参考文献の自動生成を試みた。

また、研究のもう一つの柱として、診断支援システム側の研究開発も進めた。本研究の目的である疾患知識ベースの Community-Driven Curation を実現するためには、多くの臨床医に診断支援システムを日常的に利用して頂く必要がある。そのためには、診断支援システムの有用性を高めなければならない。そこで、システムの品質向上に向けて、各種操作の利便性向上や高速化を進めた。その際、医師を対象としたセミナー形式のデモを行い、利用者からの要望や問題点の指摘を収集しつつ次のデモまでにそれらの課題を解決するという研究開発のサイクルを設け、継続的なシステムの改良を図った。

これらの活動は、自然言語処理分野の研究者との共同研究として進めた。また、手動でのデータ生成が求められる箇所については、研究代表者が直接データ生成することに加えて、医療情報系の研究機関に対してデータ生成の委託を行った。

#### 4. 研究成果

今回の研究を通じて、様々な医療用情報システムを研究開発していくための鍵となる疾患知識ベースの効率的な整備手法と、疾患知識ベースを利用する診断支援システムの研究開発を行った。研究成果として、まず、この再利用可能な知識ベース、臨床検査結果データ、参考文献データを挙げることができる。また、研究期間を通じて診断支援システムに継続的な改良を行ったことにより、ユーザー評価の高いシステムを構築することができた。これらは、今後、医療用情報システム研究を発展させていく基盤として価値が大きい。

また、従来の医療用自然言語処理は、膨大なカルテから患者情報を獲得することが一つの柱であり、医学文献からの疾患情報抽出は研究の余地が多かった。そこで、今回、疾患知識ベースの整備に際して、医学文献からの疾患情報の抽出に関する様々な知見を得ることができた。まず、医学文献における疾患の呈する所見について、利用語彙や概念について整理をすることができた。また、そうした所見についての自動抽出に関する知見を得ることができた。さらに、今まで研究が進んでいなかった医学文献に含まれる臨床

検査結果表現について、分類手法と基礎的統計を示すことができた。また、臨床検査結果について、簡便で実用性の高い知識表現手法を示した。

さらに、疾患知識ベースの自動生成から派生する形で、複数の関連分野の研究成果を挙げた。まず、疾患知識ベースの収載疾患数が増すにつれて診断結果も膨大なものとなりうるため、膨大な疾患を効率的に表示する手法として疾患の階層関係を用いることで多数の疾患を効率的に提示しうることを示した。また、膨大な疾患群からの効率的な類似疾患検索手法、膨大な疾患群のそれぞれに対する参考文献の自動生成手法、複数の疾患データベースを統合するためのブリッジコーディングのそれぞれについても、研究成果を挙げつつある。

本研究では、以上のように、自然言語処理を用いた疾患知識ベースの自動生成と実用性を備えた診断支援システムの開発を行い、様々な研究成果を挙げることができた。また、疾患知識ベースの低コストな高精度化と高性能な診断支援システムの研究開発に向けた研究体制を確立することができた。今後、テストユーザーを対象にシステム品質のさらなる向上を目指すと共に、利用統計を用いた疾患知識ベースの精度改善に向けた検討を進めていきたい。

## 5. 主な発表論文等

〔雑誌論文〕 (計 0 件)

〔学会発表〕 (計 10 件)

T. Okumura and M. Tagawa, Hierarchical representation of differential diagnosis lists for clinical decision support systems, Workshop on Interdisciplinary Design of Pervasive Healthcare Applications (IDPHA2014), May 2014.

T. Okumura, E. Aramaki, and Y. Tateisi, Expression of laboratory examination results in medical literature, Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM 2014), May 2014.

T. Okumura, Y. Tateisi, and E. Aramaki, Classification and characterization of clinical finding expressions in medical literature, the IEEE International Conference on Bioinformatics and Biomedicine (BIBM2013), Dec 2013.

T. Okumura, E. Aramaki, and Y. Tateisi, Clinical vocabulary and Clinical finding concepts in medical literature, International Joint Conference on Natural Language Processing, Workshop on Natural Language Processing for Medical and Healthcare Fields, Oct 2013.

奥村 貴史, 荒牧 英治, 建石 由佳, 疾患知識ベースの community-driven curation と診断支援プラットフォーム, 第 30 回セマンティックウェブとオントロジー研究会, 2013 年 8 月.

T. Okumura and Y. Tateisi, Interpretation of laboratory examination results and their simple representation, IEEE International Symposium on Computer-Based Medical Systems (CBMS2013), June 2013.

奥村 貴史, 医療用自然言語処理の実用的課題と戦略的テストベッドとしての診断支援プラットフォーム, 言語処理学会第 19 回年次大会, 2013 年 3 月, 名古屋大学

奥村 貴史, 集合知と診断支援プラットフォーム, 東京大学 知の構造化センターシンポジウム 2013, 2013 年 1 月, 東京大学.

T. Okumura, and Y. Tateishi, A lightweight approach for extracting disease-symptom relation with MetaMap toward automated generation of disease knowledge base, 1st International Conference on Health Information Science (HIS 2012), Apr 2012.

Y. Tateisi and T. Okumura, A Simple Approach to NTCIR-10 MedNLP task, 10th NTCIR Conference, MEDNLP Track, June 2013.

〔図書〕 (計 0 件)

〔産業財産権〕

○ 出願状況 (計 0 件)

○ 取得状況 (計 0 件)

〔その他〕

<http://mednlp.m-crisis.wide.ad.jp>

## 6. 研究組織

### (1) 研究代表者

奥村 貴史 (OKUMURA, Takashi)

国立保健医療科学院 研究情報支援研究センター 特命上席主任研究官

研究者番号 : 50553400