(別添 3)

平成 28 年度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 基盤構築研究事業)

カルテ情報の自動構造化システムと疾患数理モデルの逐次的構築に関する研究

電子カルテは患者情報が全て記録されているものの，非文法的かつ断片化した表現が多く自然言語処理を応用した利活用は困難であった．これを二次利用するため申請者等（申請者荒牧及び分担者河添が所属する研究室主宰者の大江ら）は，2008年から電子カルテから医療用語の自動抽出及び自動コーディングを行う研究に従事してきた．その成果は，日本内科学会の症例報告検索システムなどとして実用化され，現在も用いられている．本研究は，電子カルテの二次利用のさらなる実用化に向けて問題となる次の２つの課題を解決する．

（課題１）実用化可能な解析精度の達成（マッピング精度80%）

（課題２）電子カルテに組み込み可能な実装の開発

荒牧英治 奈良先端科学技術大学院大学 研究推進機構

若宮翔子（奈良先端科学技術大学院大学 研究推進機構･博士研究員）

河添悦昌（東京大学医学部附属病院 企画情報運営部･講師）

A. 研究目的

本研究の目的は，電子カルテに自由記載された文章を対象に，これを二次利用可能な状態に自動変換する技術を確立することである．これを実現するために，（問題１）現状の解析システムの解析精度を向上させ，これを（問題２）多様かつ複雑な電子カルテシステムに組み込む．これは３つのモジュールから構成される（**図1**）．

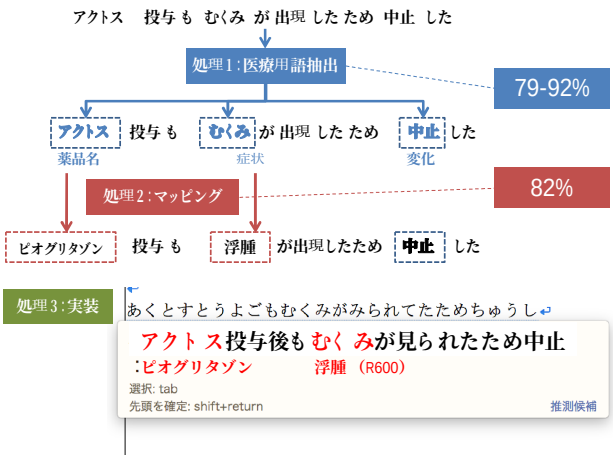

図 1: 本研究にて開発する３つの処理

（２つの言語処理技術とその実装）

（１）【処理１：医療用語抽出】

電子カルテ中の自然文から医療表現（時間表現と疾患／症状表現）を抽出する．

（２）【処理２：標準化変換（マッピング）】

自由記載された病名をICD10コードへマッピングする．これまで，出現頻度が低い（まれな）コードへのマッピングは困難であったが，前段（医療用語抽出モジュール）の結果を用いて，どのような患者がどのようなコードを付与されやすいかという確率モデルを構築する．

（３）【処理３：実装】

処理１と処理２により，既存の電子カルテ情報については後ろ向き解析が可能となるが，それでも一定の誤りが含まれてしまう．そこで，新たに電子カルテに医師等が入力する際に，標準化を行った結果をサジェストするという前向き処理機構を開発する．これにより，現場の医師の負担となることなく，自然と標準的なデータが蓄積されることを目指す．

【特色と独創性】

従来の多くの同様の研究は，ラボレベルの実験にとどまるものが多かった．本研究は新しく入力する際にも利用可能な実装も含めて開発の範疇に入れている．このような出口を持つことで，さらにデータの蓄積が爆発的に増加することが考えられ，それが解析精度を向上させるというプラスの循環を作ることができる．

B．研究方法と成果

以下の３つのシステム，リソース，アプリの開発を行い目的を達成した．

（１） 汎用病名抽出器 MedEX/J の開発／配布／評価

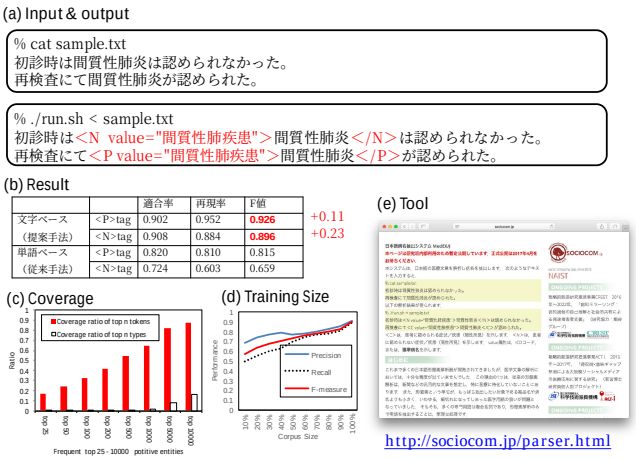

図 1: MedEX/J の概要．（a）MedEX/J への入力（上）と出力（下）．陽性所見が＜P＞，陰性所見が＜N＞としてタグ付けされて出力される．（b）タグ付けの精度．陽性所見は 0.926, 陰性所見 0.896 と，従来の単語ベースの手法を 10 ポイント程度上回っている．（c）コーパスにおける病名の出現統計．頻出する n 病名が対象とするコーパス全体に出現する病名をどの程度カバーしているかを

示したもの（塗りつぶし＝延べ単語数，白色＝種類数）．頻出する 10000 病名が 90%近くをカバーしているが，種類数としては 20%に満たない．すなわち，少数しか出現しない病名が無数にあることを示している．(d) 教師データの量（X 軸）と病名特定精度（Y 軸）．教師データが増えると順調に抽出精度は向上する．来年度増強するコーパスで 95%を達成する試算である．(e) MedEX/J 配布サイト．本ページにてツールを試験公開している．

本システムは，日本語の医療文章を解析し病名を抽出する．例えば，図 1（a）上のテキストを入力すると，図 1（a）下の解析結果が得られる．ここで，＜P＞は，患者に認められる症状／疾患（陽性所見）を示し，＜N＞は，患者に認められない症状／疾患（陰性所見）を示す．value 属性は標準病名を示す．

予備実験の結果，病名抽出においては形態素解析を用いず，いきなり文字そのものを処理する方式の方が高精度であることが分かり（図 1（b）：陽性抽出の F 値 0.926，陰性抽出の F 値 0.896），この結果を受けて，形態素解析部を省くことで，よりコンパクトな解析器を構築できることになった．速度についても汎用機（core-i7-6core 3.4GHz; Memory 32GB；70 万円相当）にて，3000 退院サマリを 120 秒で処理可能など十分実用に耐えうる．

現在は，班内および共同研究者に試験配布を行っている（図 1（e）http://sociocom.jp/parser.html ）．

（2）MedEX/J に利用する辞書「万病辞書」構築

カルテ文章調査の結果，延べ45万症状表現（種類数としては6.2万種類）が得られ，その28.3%（種類数としては87.5%）が，標準病名でカバーされていないことが分かった．このうち高頻度（頻度30回出現の5,600病名）を扱い医療従事者3名によりコーディングを行い，意見が食い違ったものはその曖昧性も残したまま辞書リソース化した（通称「万病辞書」）．この万病辞書により，現在すでにカルテに出現する80%（ただし種類数としては20%）の症状／病名を標準病名に変換可能である．

（3）日本語入力パレットの開発

日本語入力パレット（通常のIMEを用いて入力を行うと標準病名に変換した結果がサジェストされる）を開発した（図2）．

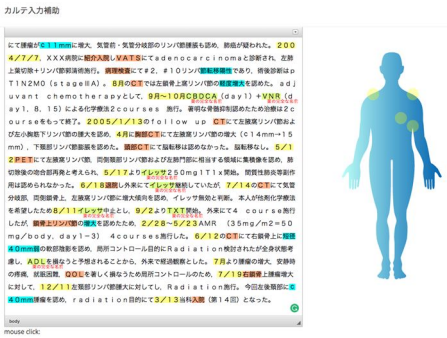

図 2:試作した日本語入力パレット．電子カルテに入力する前に，本パレット上で入力を行い，入力した病名が

標準的かどうかを確認しながら入力ができる．また，クリックにより，サジェストされる標準病名と置換可能である．

D．結論

これまで多くの日本語形態素解析器(mecab, juman など)が開発されてきたが, 医学文章の解析においては, 十分な精度が出ていなかった．この理由の１つは，従来の形態素解析は，新聞などの汎用的な文章を想定し，特に医療に特化していないことにある． また，形態素という単位が，もっぱら抽出したい対象である薬品名や病名よりも小さく，いわゆる，細切れになってしまう問題もある．

このような問題を解決するために，本研究班で開発する MedEX/J は，形態素ではなく，病名用語抽出に特化し，その後処理として，標準病名への標準化，事実性判定など，研究，臨床的に重要な処理も組み込んだ．

E．研究発表

1. 論文発表

- E.Aramaki, K.Yano, S.Wakamiya: MedEx/J: A One-scan Simple and Fast NLP tool for Japanese Clinical Texts, MedInfo, 2017.（採択）

2. 学会発表

- 矢野憲，伊藤薫，若宮翔子，荒牧英治: 深層学習による医療テキストからの固有表現抽出器の開発とその性能評価： 人工知能学会全国大会（JSAI）, 2017.（査読なし）
- 矢野憲，若宮翔子，荒牧英治： 医療テキスト解析のための事実性判定と融合した固有表現認識器，言語処理学会年次大会，2017.（査読なし）

F. 健康危険情報

該当なし