

一般口演

一般口演21

医療データ分析9（機械学習・テキストマイニング2）

2018年11月25日(日) 10:40 ～ 12:10 D会場 (4F 413+414)

[4-D-2-6] 電子カルテ診療録に対する機械学習を用いた重要キーワード抽出機能の開発と今後の診療支援機能への展望

○早川 龍雄¹, 玉田 嘉紀², 内野 詠一郎², 岩本 和樹¹, 井上 和仁¹, 奥野 恭史³ (1.富士通株式会社 第二ヘルスケアソリューション事業本部, 2.京都大学大学院医学研究科 医療情報AIシステム学講座, 3.京都大学大学院医学研究科)

【目的】急激な科学技術の発展で、医療現場もデジタル化が進み、多種多様な医療情報（リアルワールドデータ）が蓄積されている。これらのデータと、革新的に進化しているAIを組み合わせることで、患者に最適な医療を提供する次世代の高度医療が強く期待されている。高度な判断が可能なAIの構築のためには、詳細で時系列にわたる臨床情報を活用する必要があるが、これらの情報は文章記載などの構造化されていない部分も多く、電子カルテ内のテキストや数値、医療論文での記述、遺伝子情報、診断画像といった多種多様な膨大なデータを統合・連携して実践応用するためのAI技術はいまだ確立されていない。【方法】フリーテキストで記載される電子カルテ診療録は、診療ガイドラインのような整理された自然文ではなく、文字と数値の組合せや、口語表現が多いだけでなく、さらに簡略語や誤字など曖昧性を含んだ内容であることが多く、従来の機械学習による言葉の出現度や前後の文脈のベクトル抽出による単語の意味抽出や文書の類似性把握を行うことは困難であった。そこで、電子カルテ診療録そのものの構造パターンを機械学習するだけでなく、診療ガイドラインや各種医療辞書から抽出した独自のナレッジグラフを組合せ、電子カルテから重要キーワードと関連用語を抽出し、意味づけを行う技術を研究・開発している。【結果・考察】電子カルテ診療録において時系列化されたデータとSOAPの関連性を考慮しながら重要キーワードを抽出・構造化し、大量データを学習することで、電子カルテのフェノタイピングを実現することが可能となる。その要素技術を応用し、今後、診療計画のサジェッションや様々なレポートの自動作成などに利用可能と考える。今回研究・開発するAI機能が医療現場の医師の作業負および患者の経済・時間的負担の軽減に繋がることが期待される。

電子カルテ診療録に対する機械学習を用いた 重要キーワード抽出機能の開発と今後の診療支援機能への展望

早川 龍雄^{*1}、井上 和仁^{*1}、
玉田 嘉紀^{*2}、内野 詠一郎^{*2}、奥野 恭史^{*2}
^{*1} 富士通株式会社 ^{*2} 京都大学大学院医学研究科

Development of important keyword extraction function using machine learning for text data of electronic health records and perspective for the application to diagnostic aid system.

Hayakawa Tatsuo^{*1}, Inoue Kazuhito^{*1},
Tamada Yoshinori^{*2}, Uchino Eiichiro^{*2}, Okuno Yasushi^{*2}
^{*1} FUJITSU Ltd ^{*2} Graduate School of Medicine, Kyoto University

As a result of the advances in digital technology and its application to the medical field, a wide variety of medical information, a kind of real world data, has been accumulated. The emerging artificial intelligence (AI) technology is expected to analyze these data and realize advanced medical care of the next generation. Currently, a lot of patient information including records of time-series care process is stored as unstructured free text, which makes it difficult for current AI to handle these data properly and perform advanced judgment.

We are developing technologies to extract and label important keywords and related terms from electronic health record (EHR). In addition to machine learning of the structural pattern of the records in EHR itself, we combined knowledge graphs developed from clinical practice guidelines and various medical dictionaries. Our results suggested that these technologies may be useful for developing a clinical decision system which automatically suggests various reports of the diagnosis or treatment plans in the future. We expect that these AI technologies lead to the reduction of workload of doctors in the medical field and economic or time burden of patients.

Keywords: EHR (Electronic Health Records), NLP (Natural Language Processing), AI

1. 結論

電子カルテ等に蓄積された医療情報データを利活用しながら健康長寿社会実現に向け、先制医療や個別化医療へと繋がる基礎研究が大学機関で進められている。

今後の日本の先進医療に向け、より複雑化・高度化しつつある膨大なこれらの情報への対応として ICT を活用したデータ利活用の効率化が望まれている。医療情報の組合せにおける医師の暗黙知をどのように可視化とし、テキストで書かれた診療録からどのように情報を判断するかについての研究およびシステム化はあまり進んでいない。

医療機関のデータは常に蓄積されるビッグデータ(リアルワールドデータ)となっているが、同時にそれらのデータを利活用するアプリケーションが必要となってくる。IT 企業としてそのような医療機関のデータを活用するための技術の研究を同時に進めることは重要な役割であると考えられる。

著者らは、電子カルテシステム開発の知識を活用しながら、アカデミアと共同で電子カルテ診療録のテキストデータの解析の為に研究を行っている。京都大学と富士通は、新たな診療支援や創薬の実現など AI を活用した次世代の高度医療化に向けた研究開発を目指し、共同研究講座を設立した[京都大学医学部附属病院 医の倫理委員会 承認番号: R1408] [1]。研究で開発した要素技術や知見を応用にも進め、社会に成果を広く還元していくことを目指している。キーワード抽出や特徴探索、時系列での文書要約などの技術の組合せにより、医療現場の先進医療の実現に向けた診断支援システム化を検証している。大学の臨床医のもとで大学研究員がデータ解析を行いながら課題を抽出し、要素技術の開発と検証を進めている。この共同研究における要素技術のプロトタイプの開発と課題について以下に論じる。

2. 目的

電子カルテの診療録は、主に S(患者の主訴)・O(医師の所見)・A(医師の診断)・P(治療の計画)のように入力されている場合が多い。実際のカルテテキスト内容としての標準化が行われていないことが多い。診療録記載の中には、口語表現的な内容や数値データの列挙、病名や薬品名の略語などが含まれている。データは患者ごとに時系列で管理されており、何らかの治療行為に対する効果(outcome)は記載方法として統一されていない。電子カルテにおいて患者の経過の変化に関する記載や検査結果値の変化などから判断する必要があり、機械可読なデータでないものが多い。

筆者らは以前、辞書ベースによる電子カルテの時系列解析と全文解析システムの研究開発を実施した[2]。当時の辞書ベースでの形態素解析による解析の場合、以下のような課題があった。

- 常に蓄積される電子カルテ診療録内で新たに発生する単語(誤字を含む)の辞書登録では対応できない
- 抽出した単語の意味(病名・薬品名・症状・処置名など)が判断できずに単なる表記上の文字でしか判断できない

全文検索機能は、患者の横串検索や単語の出現頻度に関する解析においてランキングや単語の出現場所の把握という使い方では効果がある。患者の症状の経過や治療パターンの抽出においては、患者毎の治療効果を見る場合は単語に意味を持たせ、その出現傾向の変化と治療イベントの相関の特徴性を解析する機能が求められる。これらの重要な技術要素を実現する為、辞書による形態素解析以外に、自然文解析における出現する言葉と言葉の距離や文脈の類似度などを含めて検討する必要性を感じていた。

電子カルテ診療録で使われる単語には、その単語だけでは意味を持たない使い方が多い。例えば「痛みがない」「感覚がない」の2つの否定的文書の場合、「痛みがない」は正常な状態であり、「感覚がない」は異常な状態であるというように言葉の組合せによって意味を持つ。その為、単語にカテゴリとしての意味付けを行い、対象となる単語の前後の出現単語との組合せを文脈から判断する仕組みは重要である。

電子カルテ診療録では時系列で症状の変化が言葉と数字で併記されることが多い。言葉では状態の回復・悪化が判断できる内容でも、数値のみでは、基準値や医師の判断基準がないと状態を判断できない場合がある。

これらの問題を解決する為には、電子カルテ診療録の中に出現する単語の組合せがどのような意味を持ち、どのように変化していくのかの特徴と取得し、大量のデータから重要なキーワードを含む文章に絞っていく作業を自動化することがシステム化の課題と考えた。医療現場の医師が電子カルテ診療録および関連する医療情報(検査結果等)を利用して患者の状態を簡易に把握し、治療の計画や症状のサマリ化を支援する仕組みにつながる要素技術の研究を目的とした。電子カルテ診療録のテキストデータの特徴抽出と機械学習技術を用いたプロトタイプを開発した。

3. 各要素技術プロトタイプの概要

3.1 電子カルテ診療録での単語の意味付け

3.1.1 方法

電子カルテ診療録内の出現単語に対してタグ付けを行った学習データセットを作成し、機械学習により単語の出現パターンと前後の文脈解釈により自動的に意味付け(病名・薬品名・症状など)を実施した。学習データセットには事前に意味付けのタグを図1のように人の手で作成した。

電子カルテ診療録のサンプル文書を200文例と、CKD(慢性腎臓病)の診療ガイドライン300行程度を抜粋して意味をタグ付けし、機械学習を実施した。機械学習は、富士通が提供するAI技術プラットフォームZinrai[3]の自然文解析APIを利用した。学習済みモデルを電子カルテ診療録のサンプル1000文書に適用した。プロトタイプの検証では、多様な単語のパターンが出現する電子カルテでのフリーテキスト記載において、毎回辞書作成を行わない前提でのシステム利用を想定し、医療専門用語辞書は使わない方法での検証を行った。

3.1.2 結果

学習済みモデルを利用して、学習データセットとは別の50文例でタグにより意味付けされたデータを取得し、人の手でタグに対する○×方法で図2のようにマッチングしたところ正解

2002年に米国腎臓財団(NKF)が「CKD」慢性腎臓病(Disease) (CKD) の概念を日本腎臓学会は社会に開かれた腎臓学会を目指しており、重点課題として「Disease」まず、日本人に適合した糸球体濾過量(GFR)推算法の作成により、わが国の「CKD」成人の8人に1人が「CKD」患者であり、「CKD」患者の数は増加傾向にある。このように多くの「CKD」患者の診療は腎臓専門医のみではできません。かかりつけ医との連携が必要となります。そこで、かかりつけ医を対象とした「CKD診療ガイド」を2007年に作成し、2009年「CKD」という概念はわかりやすくモットーに提唱され、わが国のみならず世界に於ける「CKD」の概念はより大きく予後が異なることが明らかになり、まず「CKD」患者に、アルブミン尿・蛋白尿はGFRとは独立した「CKD」の進行因子であることが腎臓病の評価法も今回改訂しました。これまでは「CKD」血清クレアチニン「CKD」値のみから「CKD」GFR「CKD」推定「CKD」わが国は高齢化社会であり、年齢を加味した「CKD」かかりつけ医から腎臓専門医への「CKD」また「CKD」食事療法「CKD」や「CKD」降圧療法「CKD」に關しての改訂により、週末を返上して本ガイドの作成にあたっていただいた今井潤裕委員長をはじめ「CKD」CKD対策が次第に実を結び、新たに「CKD」診断「CKD」を開始される「CKD」患者。しかし、さらに「CKD」診療の向上を図る必要があります。「CKD」慢性腎臓病(Disease)とは、腎臓の障害「CKD」蛋白尿「CKD」「CKD」推定「CKD」GFR「CKD」は以下の「CKD」血清クレアチニン「CKD」「CKD」下肢切断「CKD」痛など、筋肉量の極端に少ない場合には

図1 診療ガイドラインによる学習データセット作成例

率は約65%という結果であった。診療ガイドラインでタグ付けた内容の比率として、「病名」「検査」が多く、学習済みモデルで誤認識されるケースが多いことが分かった。無作為に学習データセットを多くするのではなく、意味付けするタグの出現がなるべく均一になるような文章パターンを多く作り、独自の単語グループ辞書や排他辞書を活用することで、正解率を向上させることができた。今回の検証により、電子カルテ診療録のように大量データが学習できない場合においても、意味付けされた単語を少ない労力で抽出できることが確認できた。

今後、学習データセット数を増やし、事前のデータクレンジングを実施する対策を行うことで、単語抽出精度を高めることが可能だと考える。

3.2 電子カルテ診療録文書の要約

3.2.1 方法

電子カルテ診療録で記載されている患者毎のデータのパターンを検証した。1 患者の入院期間中のサンプルの場合、

- 入院に至る記録・退院に至る経緯
- 入院時の経過記録(テキスト)
- 各種検査の記録(テキスト+数値)
- 処方・注射・処置の記録(オーダ)

が主に記載されている。bとcのデータ量はかなり多く、短時間で患者の状態を電子カルテ診療録のみで把握することは難しいことが臨床医を含む著者らの調査検討で分かった。dのデータは電子カルテシステムが記載した情報が主であり、記載内容がパターン化されていることが多い。患者の症状経過を把握する為に、電子カルテ診療録で記載されているこれらの文章を鳥瞰的に把握する仕組みが必要と考える。電子カルテ診療録の患者毎の上記のような記載全体を要約し、患者の状態を把握する目的での要素技術の検討を行なった。

電子カルテ診療録サンプルに対してクラスタリングを実施し、文書の分類化を実施した。各クラスターで出現する単語や単語と単語の関連性などから各クラスターが主にどのような特徴の話題であるかを判断する。それらクラスター内で代表文書を抽出し、3.1の意味付けられた単語から状態に関するラベルを付与することで要約文書を作成する方法で検証を行っている。

3.2.2 結果

電子カルテ診療録のサンプル文書をベクトル化し、クラスタリングした様子を多次元尺度法で可視化した結果が図3である。薬品や症状などの記載内容の特徴性である程度分類化できていることが分かった。

また、上記結果から電子カルテ診療録のサンプル文書の類似度で文書同士にリンクを張り、上位3500件の関係性を

【学習済みデータに存在する(パターン・単語も同じ)】
CKD患者はクレアチニン値が上昇する。
Disease>CKD</Disease>患者は「CKD」クレアチニン「CKD」値が「Shojo」上昇</Shojo>する。
【学習済みデータに存在する(パターン・単語は別)】
CKD患者は尿蛋白値が上昇する。
Disease>CKD</Disease>患者は「CKD」尿蛋白「CKD」値が「Shojo」上昇</Shojo>する。
【学習済みデータに存在しない(パターン・単語も別)】
肺炎患者の症状による喘息の疑いがある。
Disease>肺炎</Disease>患者の「CKD」症状「CKD」による「CKD」喘息「CKD」の「Shojo」疑い</Shojo>がある。
肺炎患者の症状による喘息の疑いがある。
Disease>肺炎</Disease>患者の「CKD」症状「CKD」による「CKD」喘息「CKD」の「Shojo」疑い</Shojo>がある。
急性腎障害(AKI)はがん治療において発生することが多く、多臓器不全の症状により死亡する場合がある。
Disease>急性腎障害</Disease>(AKI)は、がん「CKD」治療「CKD」において「CKD」発生「CKD」する「CKD」多「CKD」臓器不全「CKD」の「CKD」症状「CKD」により「CKD」死亡</CKD>する場合がある。

図2 機械学習結果から抽出したタグの正解比較例

出力したのが図4である。この図のように、複数の文書がある特徴性パターンでグルーピングされており、中心に向かって代表的な文書に収束されている構造となっている。

これから、文書のクラスタおよび類似度のグルーピング結果から最も中心に存在する文書を抽出し、要約結果として文書整形後に出力することで、要約カルテ文書として出力を行う。以上により、本技術を利用することで、電子カルテ診療録の1患者の膨大な記載内容を集約することが可能になると考える。さらにクラスタリングを可視化した結果を確認することで、「患者を層別化すること」、「電子カルテから重要な記述を抽出すること」、「症状を分類すること」など様々な技術への応用が期待できる。

3.3 電子カルテ診療録の時系列特徴解析

3.3.1 方法

電子カルテ診療録は時系列データであり、時間軸で状態が変化していく為、単語への意味付けや文書要約だけでは状態の変化を抽出することはできない。電子カルテ診療録はパターン化された構造での時系列データセットではなく、多様なテキストや数字の組合せ情報である。それらを解析可能なデータとする為に患者の診療録や検査結果などからテキストや数値データを取得してから日付毎・項目毎の時系列データとしてまとめ、期間ウィンドウ(時間幅)とスライド幅(順次スライドする時間幅)に従って前記時系列データを分割する。期間ウィンドウとスライド幅については診療録内のデータ量で事前にパラメータとして図5のように決定しておく。

分割された各期間ウィンドウ内に記載のテキスト文からの意味付けされた単語の抽出と、検査結果等からの関連する数値データの抽出を行い、スライド幅毎に分割した期間ウィンドウ毎にまとめ、それぞれの期間ウィンドウごとの単語の出現頻度の変化や数値情報の変化の相関や因果関係などの特徴性をダイナミックベイジアンネットワーク等のアルゴリズムを利用して特徴性を探索する。患者の状態を示す多様な情報は複雑な相関関係を示すことが多い。単語や数字項目を変数として相関性分析を行う場合に背景の第3因子が影響

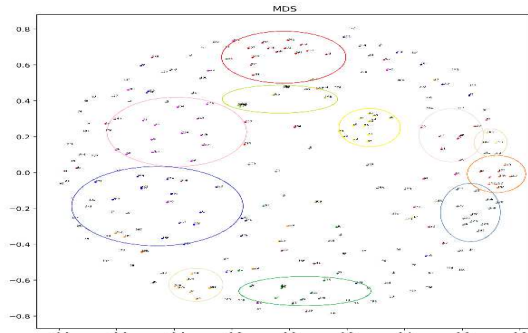


図3 電子カルテ診療録サンプル文書のクラスタリング例



図4 文書の類似度による特徴パターンの可視化例

を与える場合が多い。それら変数間の特徴を第3因子も考慮してネットワークとして特徴性探索するベイジアンネットワークを利用し解析を行う方針とした。

3.3.2 結果

図6は、時系列での0週～3週の項目間の相関性と13～16週の項目間の相関性の2つの期間ウィンドウの違いであり、特定の項目で相関性に変化が発生していることがわかる。

この変化を患者の状態の変化としてとらえ、特徴性を抽出するという技術を開発している。この技術により、一連の電子カルテ診療録の要約結果から変化点をとらえ、患者がどのような状態になっているかの解析を行うことが可能となる。【特許出願中】

文書要約技術としては、「Feature Base」「Cluster Base」「Graph Base」「Knowledge Base」などの技術もあり[4]、今後精度向上に向け、これらの技術との組合せについても検証を実施していく方針で検討している。

4 課題

3.1から3.3により、電子カルテ診療録のサンプルデータにおける特徴的な単語の意味付け、テキストデータの要約、記載内容および数値データの変化の抽出を行うそれぞれ要素技術の検証と結果取得を行った。大量の電子カルテ診療録の記載内容を鳥瞰的に把握できるような仕組みを構築することができた。

実際の電子カルテ診療録には、標準化されていない用語の記載や、口語表現化された文書、連続のネガポジ表現(「～でないわけでない」等)、文字と数字の組合せの列挙、オーダ発行時の固定文字列、テンプレートによるパターン表現などが多く、辞書を利用した形態素解析により抽出した単語での特徴性抽出や時系列変化の解析だけでは、正しく患者の状態変化をとらえることは困難である。この点は学習データセット数を多くして学習を繰り返しても精度を向上することはできないと考える。

電子カルテ診療録のテキスト解析を行う重要なポイントは、学習データセットを作成する際のデータクレンジングと辞書の整備であると考えられる。単語の意味付けにおいて、複数人で学習データセットを作成する際に、アノデータ毎に設定するタグ

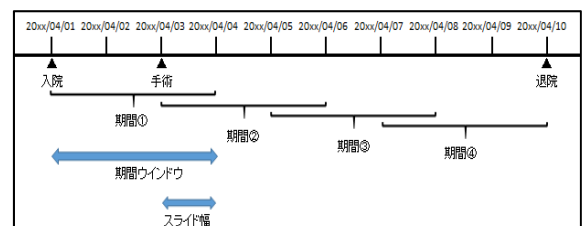


図5 期間ウィンドウとスライド幅の関係(パラメータ)例

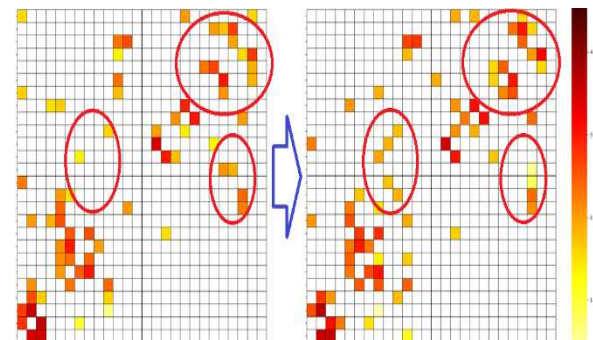


図6 期間ウィンドウ間の項目の相関の変化の可視化例

が異なることによりバイアスが発生する課題についても意味付け定義の正解率を低下させる要因になる。時系列解析においては特定の期間ウィンドウの中にデータが存在しない時系列スパースな状態であることが多く、正しく変化点を特徴性として抽出できないケースも多い。医療情報はすべての患者で一律なデータパターンを持っているわけではない為、電子カルテでの症状の同定化は困難であると考ええる。

今回のプロトタイプでは特定の疾患に関する電子カルテ診療録の学習データセットに限定して実施したため、他の疾患で同じ結果を得ることはできない。疾患別の診療ガイドラインなどを利用した解釈の追加などで医師の知見を意味付けや特徴解析の仕組みの中に加えていくことが必要である。

既存の医療ビッグデータを解析することで、データクレンジングのあり方やデータ構造化、層別化技術の高度化などが今後必要な技術であると考えられる。単なる技術的な可能性の追求ではなく、産学連携での共同研究の中でデータサイエンティスト・バイオインフォマティシヤンの解析上の構造の知識、臨床医のデータからの知見の明示化に必要な知識、IT 技術者のビッグデータ・AI 解析技術の知識を終結したデータ解析の実現が必要であると感じる。

5. 考察・結論

本論文では、機械学習技術を利用した電子カルテ診療録の解析要素技術のプロトタイプの開発により、リアルワールドデータの効率的な解析の実施について述べた。

今回構築したプロトタイプ開発では、電子カルテ医療データの集約化により、診療現場医師の作業負担軽減につながる要素技術開発の布石となる研究成果を導くことができ、同時に将来的に実臨床として利用する為の課題抽出を行うことができた。

今後、現在の研究を推進する中で課題に対する技術的対応を実施していく点と合わせ、今後の電子カルテにおける標準化・構造化を提唱していくことも重要であると考えられる。この分野の研究が進むことで、患者の主訴から診断に必要な問診・触診等の情報をレコメンドし、早期に患者の状態を把握し治療を行うような医療現場となることが求められる状態であると感じる。さらにはゲノム情報とマルチモーダルの連携することで、個別化医療の推進とともに患者の通院・治療にかかるトータルの時間・コストの軽減化が社内保障費の抑制にもつながるという点が重要である。

様々な研究により生み出されるこれら要素技術が、効果的治療法などの医師支援機能として電子カルテ機能に実装されることで、医療品質や患者の QOL (Quality Of Life) への向上に寄与できることが期待できる。今回検討電子カルテ診療録の解析における医師作業軽減化のシステム化を診療支援モデルにより、医療ビッグデータ解析を次世代電子カルテシステムの一機能として実現し、社会に還元されて患者の時間・コストの負担軽減につながるソリューションとなることを期待する。

参考文献

- 1) 富士通プレスリリース, 「京都大学と富士通、高度医療化に向けた AI 活用の共同研究講座を開設」
[<http://pr.fujitsu.com/jp/news/2018/01/24.html>]
(2018/01/24 アクセス)
- 2) 吉田茂, 早川龍雄 他, 「Hadoop を用いた電子カルテの統計分析とタイムライン検索〜次世代電子カルテ閲覧システムに向けた研究〜」, 第 32 回 医療情報学連合大会 (第 13 回 日本医療情報学会学術大会), 2012
- 3) 富士通プレスリリース, 「当社が培った AI 技術を「Human Centric AI Zinrai」として体系化」, 2015
[<http://pr.fujitsu.com/jp/news/2015/11/2.html>].
(2018/08/15 アクセス)
- 4) Yogan J. K., et al, 「Automatic Multi Document Summarization Approaches」, Journal of Computer Science, 2012