様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 １９ 日現在

機関番号：８２６０２
研究種目：基盤研究(C)（一般）
研究期間：2015～2018
課題番号：１５Ｋ０８８４５

研究課題名（和文）高品位な知識抽出を実現する三階層オントロジーフレームワークの開発

研究課題名（英文）Development of a three-tierd ontology framework to realize high-quality knowledge extraction

研究代表者
木村　映善（Kimura, Eizen）

国立保健医療科学院・その他部局等・統括研究官

研究者番号：２０３６３２４４

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：医療記録から潜在的な有害事象を発見するには、有害事象につながる原因・症状から推察できる能力を獲得する必要がある。有害事象を定義する文章から計算される概念ベクトルに似た概念ベクトルを持つ文章が有害事象を示唆しているという仮説を検証するための手法の開発を試みた。医学用語の品質のよい意味分散表現を獲得するために電子カルテの記載内容を構文解析し、意味分散表現を抽出した。周辺概念をまとめて捕捉できるように概念間関係を定義しているUMLSと医学用語間のマッピングをするために、意味分散表現ベクトルのクラスタリングによるマッピング候補の抽出手法を開発した。

研究成果の学術的意義や社会的意義
医療記録から潜在的な事象を読み取る能力を獲得することは人工知能が医師の判断能力に近づくために必要なプロセスである。良質な意味分散表現は医療記録の分析から得られること、周辺概念からの推察を実現するためにオントロジーとの連携が重要であること、質の良い意味分散表現の獲得にむけて、表記揺れの収束や異常値の検出等の追加的な処理を加えることの必要性の知見を本研究は示した。

研究成果の概要（英文）：Acquiring an ability that can infer the causes and symptoms leading to an adverse event is required to discover potential adverse events from medical records. We tried to develop the methods to test the hypothesis that a sentence with distributed representations of words in vectors similar to the vectors calculated from sentences defining adverse events may include adverse events. In order to obtain quality distributed representations of medical terms, we analyzed the contents of electronic medical records. We developed a method for extracting mapping candidates by clustering and detecting outlier of distributed representation of words in vectors to map between UMLS; the ontology defines inter-concept relationships and medical terms. Binding medical terms with the ontology is intended to enable inferring from peripheral concepts without direct concept by back-tracing peripheral concepts from hierarchical relationships of concepts.

研究分野：医療情報学

キーワード：自然文章解析　機械学習　オントロジー　ターミノロジー　意味分散表現

様式Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

国際的な臨床試験や症例登録においては、SNOMED-CT 等の国際医療用語集の規則と構成内容に照らし合わせて登録データを作成する必要があるため、単なるデータソースからのデータ転送とマッピングのみでは解決できない。また、日々の診療において、医師は患者の観察や検査結果などを総合し、高度な医療ドメインの背景知識に基づいて推論している。しかし、その過程において認識されたリスク要因は、医療従事者であれば当然に認識するものとして、あらためて診療記録に記載しないことがある。結果として、危機管理情報が共有されず、後日の医療インシンデントの誘因となったことが指摘されることが多々発生している。このように、多忙な医療現場において、必要最低限の記述に留められた診療記録から、人手を極力介することなく、自然言語処理(Natural Language Processing)によって暗黙的な情報を抽出し、一次・二次利用を実現する技術の開発に期待が持たれている。

現状は医療に特化したコーパスや医療辞書の構築によって一定精度のデータ抽出に成功しているが、あくまでも探索対象検出精度の深化や「ほどほどの」情報要約にとどまり、臨床概念を理解した臨床判断支援のロジックや、医師に変わるリコーディングを構成するレベルには到達していない。

国際的な研究のデータ登録には、SNOMED-CT 及び使用する各国語にローカライズされたターミノロジが利用されることが多い。しかしながら、SNOMED-CT は概念を多重継承する設計をとっており、疾患概念と異常状態が明確に分離され正確に推論できない危険性を有する(Schulz S,et al.Stud Health Technol Inform.)。従って、一次情報源となる診療録において、診療録に記載されている文章を構文解析し、それらの単語を単純に SNOMED-CT に置き換えて知識表現の抽出の足場とすることは問題を孕んでいる。一方で、有用な臨床判断支援を構築するには、疾患概念の理解にもとづいたコンテキスト下での病態の変化の理解が不可欠であるとされている。古崎らは、原因と途中経過を含めた一連の状態変化の連鎖とそれにより引き起こされている結果状態の総体として疾患を捉えることにより、この課題の解決を試み、さらに今井らによって糖尿病患者の状態を検査値から特定し、病態の親展・合併症の発症リスクを予測するモデルの開発の試みがなされている。しかしながら、検査値等定量的に捕捉できるデータではなく、診療録や問診にみられる自然文章からの NLP による知見の抽出可能性は現在も未知数であり、またその前段階の準備として診療情報中に現れる用語文字列を診療録から収集し、オントロジーの概念との対応関係を付与していく作業が必要である(今井健,言語・音声理解と対話処理研究会 61,2001)。

以上の学術的背景を踏まえて、オントロジーと診療録から抽出した用語とを対応づけ、オントロジーによる知識推論を実現することで、従来よりも高品位な医療知識抽出を実現できるシステムの実現可能性を検証することを着想した。これらの目標を念頭に、まずは具体的な知識抽出タスクを想定し、そのタスクを実現するための要素技術の検討を試みることとした。

２．研究の目的

具体的な知識抽出タスクとして、医療記録に直接書かれていない有害事象を検出することを設定した。潜在的な有害事象を発見するには、有害事象につながる原因・症状の候補をみつけ、医学的知識を背景として推論する能力を獲得する必要がある。有害事象につながる原因・症状の候補をみつけるにあたり、有害事象を定義する文章から計算される概念ベクトルに似た概念ベクトルを持つ文章が有害事象を示唆しているという仮説を立てた。この仮説について下記の 2 つの問題意識を持ち、本研究期間では主に(1)の課題に取り組んだ。

1)データの意味分散表現、コンテキストを勘案して正確な概念・用語のマッピングが可能か。

知見を抽出しても、そこから臨床研究や症例登録等の二次利用につなげることとは別である。独自オントロジーによる推論を通して得た知見を、さらに提出先が要求しているオントロジーとコンテキストに照らし合わせてマッピングを行う必要がある。そこで、意味分散表現を利用して、概念マッピングを行う方法を検討してマッピングに関する知見を蓄積したい。

2)NLP による処理で直接記述がされていない事象についても、潜在的な知見として抽出が可能か。

先述した通り、医療従事者にとって当時のデータ・状況により自明的に把握されるものについては記載が省かれることが多い。明示的に記載がなされていないものであっても、オントロジーを利用して当時の記述や結果から強く推論されるものを候補として提示させる。検証シナリオとして、アレルギーや有害事象の可能性があるものの検出を試みる。

３．研究の方法

診療禄に記載されている文章および、その文章を構成している単語群は、医学分野の概念を表すために使用されているので、通常のインターネット上で入手できる文章よりも、より的確な医学的概念に対応する意味分散ベクトルを抽出できるという仮説を立てた。電子カルテシステムに記載された診療録の SOAP 欄から文章を収集し、医学用語辞書と comejisyo と汎用的な大規模辞書 mecab-ipadic-neologd の 2 つをユーザ辞書として再構成し、mecab で分かち書きを生成する。英語のエントリーは米国医学図書館から提供されている specialt lexicon ツールを利用して正規化した。日本語のエントリーは、neologd で利用されている正規化ルーチンを利用した。UMLS における概念間の階層構造(MRREL テーブルにて定義)の中で末端に位置(当概念において子となる概念関係(chd:child)がないもの)する概念であり、かつ親関係(par:parent)、兄弟関係(sib:sibling)の関係性の定義数が多い上位群から無作為に検証候補となる概念を選択する。
その検証候補としての概念に紐付けられている用語が正規化されたもの(nstr:normalizedstring)を mrxns_eng テーブルから抽出し、前述で構築した英和辞書の正規化された英語のエントリーとマッチングし、当該エントリーに対応する日本語を訳語の候補として抽出する。このような抽出をした理由は、末端の方がより具体的な概念であり、特定の用語と対応づけやすいことと、親・兄弟の関係性が多いことにより同義語から多くの翻訳候補が出現し、後述する外れ値の算出評価によいサンプルとなると判断したためである。解析環境として anaconda4.3.0,python の機械学習ライブラリ sklearn を使用した。単語の分散表現は fasttext を利用して学習した。機械学習ライブラリ gensim から fasttext を呼び出して、前項で処理した分かち書き済みファイルを 300 次元の skipgram モデル、コンテキストウィンドウサイズは 8 で学習させた。訳語の候補として抽出された単語について、fasttext で抽出した分散表現としてのベクトルを抽出する。候補単語数が n 個とすると、300 次元のベクトルが n 個ある集合として定義される。この集合の中から外れ値を検出して、外れ値に対応する単語を訳語の候補から除外する。外れ値検出アルゴリズムとして sklearn で実装されている One Class SVM、Elliptic envelope、Isolation forest、Local outlier factor を使用した。各分類器の初期パラメータは sklern で設定されているデフォルト値を適用した。

４．研究成果

電子カルテシステムから抽出されたデータは、延べ 10325079 行、4016152166 単語より構成される分かち書きデータとなった。全体でユニークな単語数は 371684 となった。ランダムに抽出された 10 個の UMLS の概念について検証した。各々の UMLS 概念の訳語候補として定時された単語群に関して、分類器が外れ値として検出したものについて、人間が判断したものとの合致を調べたところ、全体的に One Class SVM が優位な結果を出していたが、全体的にスコアが低い Isolation Forest で precision が最高スコアを出すなど、分類器の特徴が強くでる結果となった。Isolation Forest の仮定は「正常値は数が多く相互に近似していること、外れ値は数が少なくて相互に異なること」としている。外れ値となるような単語が少なく、外れ値に対応する単語が外の単語から離散している場合は Isolation Forest による適合率は高まることが考えられる。しかし、再現率が低いので、他の分類器と比較して外れ値に対して緩やかな基準で取りこぼしている可能性がある。英和の自動マッピングにあたっては、本来訳語の候補とすべきものを外れ値として扱わないような安全側に倒れた判断をし、最終的な判断を人間に仰ぐことが望ましいと思われる。すなわち、今後の実装を進めるにあたり、全般的に高評価であった One Class SVM を主な分類器として採用し、One Class SVM で外れ値として検出されても、Isolation Forest による判定と対立すれば、外れ値としない処理を加えることを検討する。もしくは、複数の学習器を使うアンサンブル学習での検証を進めることが必要であろう。外れ値の候補が増えるが、今後、オントロジーの概念の階層を配慮して絞り込む処理によって、外れ値の境界上の評価にあったものがふるい落とされる可能性があり、最終的には外れ値が単純に増加するのではないとみている。

表記揺れによる影響の一例として、核レンズ変性症の表記揺れ・同義語について、比較的正確に正常範囲とみなしているが、「同値性」「関連性」「水曜日」「手関節離断」等、訳語として認めるべきではないものも混同していた事例があった。略語を中心とした、他の概念でも使われている用語を概念の英語の表記へのマッピングの過程で引きずり込んでしまっているものである。「wd」「wnd」は Wilson Disease の略語として使われるであろうが、他の意味にも捉えられる。意味が違うものが入るので、他の正当な単語のベクトルから距離が離れそうなものであるが、「水曜日」が外れ値として分類されていない。原因を調べたところ、大元の記載の中に肝レンズ核変性症に関する記述と、水曜日の記述が近接している箇所があり、意味分散表現上は近い距離として評価されためと思われる。約 40 億単語より構成されている医療情報であっても、このような意図しない意味分散表現上の近接性を排除するのには不十分な可能性がある。より正確な分類を実現するためには、より多くのデータが必要なことを傍証していると考えられる。医学用語は複数の単語の組合せからなる用語も多いため、構成要素となる単語単位で分かち書

きしないように、固有表現を多く網羅するべくユーザ辞書として利用した。しかし、fasttext における sub word information 手法に、敢えて意味のある範囲で最小限の長さを持つ単語のみで分かち書きするようにした方が、より実際の概念の関係・操作の結果を提供する意味分散表現を獲得できる可能性も考えられる。以上の研究については文献[8]にて詳細に確認できる。

また、本研究に派生して、様々な取り組みをした。診療記録の解析にあたり、有害事象の手がかりとなるアレルギー情報に関する調査[11]、患者基本情報の状況[1,6,10,12,15]、データの二次利用と標準化について FHIR を中心に調査と実装の検証を行った[3,4,7,13,14,16,17,18]。

５．主な発表論文等

〔学会発表〕（計 21 件）

1. 栗原幸男,石田博,木村映善,近藤博史.臨床意思決定支援の要としての患者プロファイル情報（PPI）を考える.医療情報学 38(Suppl).2018:265-7.
2. 木村映善.RealWorldData を活用する観察研究データベースの考察.保健医療科学.2018;67(2):179-90.
3. 木村映善,山本景一.EDC への電子カルテからのデータ取り込みの標準化に関わる取り組み.医療情報学 38(Suppl).2018:40-4.
4. 木村映善.SS-MIX への FHIRWeb サービス実装の試み.医療情報学 38(Suppl).2018:1074-6.
5. 栗原幸男,石田博,樫部公一,木村映善,島井健一郎,田中武志,etal.患者プロファイル情報管理の課題と改善策.医療情報学.2017;37(3):125-33.
6. 栗原幸男,石田博,木村映善,近藤博史,島井健一郎,田中武志,etal.実態調査に基づく患者プロファイル情報の実用的な標準規格の検討.医療情報学 37(Suppl).2017:462-5.
7. EizenKimura,KeiichiYamamoto.MappingdataitemsbetweenEDCandEMRusingFHIRterminologyservice.2018JointSummitsonTranslationalScience.2018:523.
8. 木村映善,蒲生祥子,石原謙.分散意味表現を利用した UMLS の概念と日本語の医学用語間のマッピングの試み.医療情報学 37(Suppl).2017:1181-5.
9. 木村映善,山本景一.CDISC/ODM の概念マッピングによる EDC と EMR 連携の試み.医療情報学 37(Suppl).2017:159-63.
10. YukioKurihara,HakuIshida,EizenKIMURA,AkiraGochi,HiroshiKondoh,Ken-ichiroSHIMAI,etal.TheinequalityofpatientprofileinformationinJapanesehospitalsStudHealthTechnolInform.2016;228:412-5.
11. 蒲生祥子,木村映善,石原謙.患者基本プロファイルのアレルギー項目の用語集編成に関する検討.医療情報学 36(Suppl);2016/11/232016.p.846-8.
12. 栗原幸男,石田博,木村映善,近藤博史,島井健一郎,田中武志,etal.病院情報システムにおける患者プロファイル情報項目の保有状況調査.医療情報学 36(Suppl).2016:1086-8.
13. 高月常光,崔ムン相,張彰祐,木村映善,蒲生祥子,石原謙,etal.パブリッククラウドと SDM を利用した２次利用環境の構築.医療情報学 36(Suppl);2016/11/232016.p.168-70.
14. 木村映善,高月常光,崔ムン相,張彰祐,蒲生祥子,石原謙,etal.標準化されたデータソースを用いた SDM 構築.医療情報学 36(Suppl);2016/11/232016.p.160-1.
15. 栗原幸男,近藤博史,入江真行,木村映善,合地明,高井康平,etal.SS-MIX2 ベースの地域医療連携システムにおける患者プロファイル情報統合の課題.医療情報学 36(Suppl).2016:238-9.
16. 木村映善,石原謙.ArdenSyntax と FHIR を利用した臨床判断支援ロジック記述環境の開発の試み.医療情報学.2016;35(6):283-96.
17. EizenKimura,KenIshihara.InternalDomain-SpecificLanguageBasedonArdenSyntaxandFHIR.StudiesinHealthTechnologyandInformatics;2015/08/212015.p.955.
18. EizenKimura,KenIshihara.VirtualFileSystemonNoSQLforProcessingHighVolumesofHL7Messages.StudiesinHealthTechnologyandInformatics.2015;210:687-91.
19. 木村映善.医療分野への人工知能適用に関する研究のトピックス.愛媛医学.2015;34(4):197-202.
20. 木村映善,石原謙.FHIRTerminologyService の実装と評価.医療情報学 35(Suppl).2015:910-1.
21. 木村映善,蒲生祥子,石原謙.StreamComputing を医療情報システムに導入する試み.第 43 回日本 M テクノロジー学会大会講演論文集.2016:11-6.

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究分担者
研究分担者氏名：岡本和也
ローマ字氏名：OkamotoKazuya
所属研究機関名：京都大学
部局名：医学(系)研究科(研究院)
職名：准教授
研究者番号（8 桁）：60565018

研究分担者氏名：今井健
ローマ字氏名：ImaiTakeshi
所属研究機関名：東京大学
部局名：医学(系)研究科(研究院)
職名：准教授
研究者番号（8 桁）：90401075