平成 29 年度厚生労働省科学研究補助金
（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

自然言語解析による診断名判断システムの開発

研究分担者　宍戸　稔聡　国立循環器病研究センター・研究推進支援部・部長
平松　治彦　国立循環器病研究センター・情報統括部・部長
上村　幸司　国立循環器病研究センター・研究推進支援部・室長
竹村　匡正　兵庫県立大学大学院・応用情報科学研究科・教授

研究要旨

診断名の判断には、電子カルテの蓄積データ、特に、自然文で記載されるカルテ記事が重要である。また、カルテ記事の分析においては、関連する情報を利用することで分析精度の向上や効率化が図れると考えている。そこで、電子カルテシステム内のデータの所在や形式について調査し、文字コードの問題があったが、必要な記述情報を抽出することができた。今後、必要な情報を簡便に抽出できる仕組みとして、基幹システムや部門システムのデータを集約・管理できる統合 DB の開発を行う予定である。また、データ抽出の際、適切な匿名化、暗号化を適用するための仕組みと匿名化のための辞書の整備も検討する必要があることがわかった。

次に、電子カルテシステム内に蓄積された所見等の記述情報から、自然言語処理および機械学習を用いて、「症状記載」と「その他の記載」に分類する自動判別器を作成した。得られた自動判別器は、正答率 80%以上、感度・特異度ともに 70～80%の性能を有していた。今後は、より多くの高精度な教師データの作成と、いくつかの機械学習手法による比較評価および最適な手法の検討を行っていく。

A．研究目的

本研究の目的は、電子カルテシステム内に蓄積された所見・報告書・サマリなどの記述情報から、自然言語処理および機械学習を用いて診断名を推測、判断できる枠組みを構築することにある。

しかし、対象とする記述情報は、様々なシステムで作成・保管され、形式も多様であるため、本研究で利用できる形式で抽出・収集可能か確認が必要である。

また、カルテに記載されている内容は既往歴・主訴・現病歴・症状・検査結果・評価・治療方法（処置・処方・手術など）・経過など様々である。その中には、病名・治療方法などの医学用語、日付・検査値などの数値情報、人名などの個人情報等、種々の情報が混在しており、自然言語処理による形態素解析やそれに伴う機械学習の精度の向上には多くの問題が想定される。

そこで、今回、第１段階として、電子カルテシステムに保管されている記述情報について所在や形式を確認し、抽出されたデータついて検証を行い、自然言語処理や機械学習を行うことが可能か評価を行った。

次に、電子カルテシステムに記載された記述情報に対する自然言語処理および機械学習の有用性を評価するために、対象を電子カルテシステムの SOAP 記載のうち S と O の記述情報に絞り、有害事象などの把握につながる「症状記載」について判別・予測できるか検証を行った.

本稿では、以上の 2 点について報告する。

B．研究方法

1．電子カルテシステム内のデータ確認

当センター電子カルテシステム（NEC MegaOak HR）における記述情報（SOAP，退院サマリなど）の抽出を行うためにデータベースのテーブル構造の調査を行った。次に、調査結果に基づきデータの抽出を行い、抽出された記述情報の内容を確認した。

2．機械学習を用いた症状記載の自動抽出に関する検討

臨床研究業務担当者（製薬企業薬剤師）1 名が、10000 件の電子カルテ記述情報（S と O）

を対象に、有害事象に関連する「症状記載」データを手動で抽出し、「症状記載」の有無を判別し教師データを作成した。

次に、上述の教師データに対して自然言語処理を用いて形態素解析を行い、カルテ記述情報の形態素（単語）を抽出した。形態素解析は、医療用語辞書（30 万語）を用いる場合と用いない場合で行った。

最後に、文章内に出現した各形態素（単語）を 1 次元とした線形サポートベクターマシンを用いて学習させ、カルテ記述における「症状記載」と「その他の記載」の自動判別器を作成した。

作成した自動判別機の性能は、10 分割交差検定を用いて評価した。

（倫理面への配慮）

本研究において診療データを利用する際には、国立循環器病研究センターなど参加施設の倫理委員会の承認を得てその指示に従う。情報収集協力病院からデータを収集する際には、個人情報は削除して連結可能匿名化とし、個人識別情報および対応表を施設外に持ち出さないように厳格に管理する。

C．研究結果

1． 電子カルテシステム内のデータ確認

所見・報告書・サマリなどの記述情報は、テキストデータとして抽出することができた。内容を精査した結果、一部、文字コードの問題による文字化けなどが発生したが、適切に文字コード変換することで、必要な記述情報を抽出することができた。

2． 機械学習を用いた症状記載の自動抽出に関する検討

エキスパート 1 名が手動で行った「症状記載」の有無の判別結果は、S 情報：有 964 件 無 3672 件、0 情報：有 1285 件　無 4076 件だった。

自然言語処理における形態素解析の結果は、医療用語辞書を用いた場合、形態素（単語）の数は 3480 種類、辞書を用いない場合、13856 種類だった。

辞書を用いた場合と用いなかった場合の形態素（単語）を学習して作成された 2 種類の自動判別器を、10 分割交差検定で評価した。その結果、辞書を用いなかった場合のほうが多少性能はよかったが、双方とも正答率は 80%以上あり、感度・特異度ともに 70～80%の性能を有していた。

D．考察

1． 電子カルテシステム内のデータ確認

本研究の目的である非構造化データ（記述情報）の自然言語処理や機械学習をするためには、対象とする所見・報告書・サマリなどの記述情報の所在・保管形式・データ形式などを把握する必要がある。今回の結果から、様々なシステムで作成・保管され、形式も多様な記述情報が、本研究で利用できる形式で抽出・収集可能か検証することができた。

また、抽出した記述情報の中には、本人・家族・医療スタッフ等の氏名や、人物名が含まれた病名（例：橋本病）などが混在していた。しかし、人名辞書などを用いた単純なマスキングでは、適切に匿名化することはできなかったため、辞書のチューニングや手動で匿名化を行う必要があった。今後、最適な匿名化手法についての検討も必要である。

以上の結果は、他施設における情報抽出・収集においてもフィードバック可能である。そのため、多施設間で大規模なデータを収集する際には有用な知見となりうる。

2． 機械学習を用いた症状記載の自動抽出に関する検討

今回行った結果から、自然言語処理を用いた機械学習が「症状記載」などのイベント判別・予測に有用であることが示せた。

ただし、教師データはエキスパート 1 名により手動で作成されたため、主観によるバイアスが混入している可能性がある。したがって、学習の精度向上のためには、複数名による教師データの精査が必要である。

また、学習効果は学習データの数にも相関するため、質の良い学習データを可能な限り増やす必要があるが、学習のための演算時間とのトレードオフでもあるため、適当な数を評価する必要もある。

形態素解析の結果、辞書有のほうが抽出単語数は少なかった。これは、医療用語が細切れにならず適切に抽出されている可能性を示唆しているが、抽出単語の精査を行った上で、辞書のチューニングを行う必要がある。

自動判別機の性能は、辞書有と無であまり差はなかった。辞書有の場合、抽出された単語数が辞書無の 4 分の一以下だったため、学習時間を考えると、辞書有で学習データを作成する方が有用であると考えられる。

さらに、機械学習のアルゴリズムはSVM以外にも何十種類もあり、学習方法も異なる。最も優れた手法や何にでも使える手法というものはないため、適切なアルゴリズムを探すには試行錯誤に頼らざるを得ない部分がある。ただし、アルゴリズムの選択は、扱うデータのサイズや種類、データから導き出したい見解、その見解の活用方法によって決まってくる部分もあるため､先行事例を参照しながら、より最適なアルゴリズムの検討を行っていく必要がある。

E．結論

自然文で記載される電子カルテ記事の分析を行うために、電子カルテシステム内のデータの所在や形式について調査した。文字コードの問題などがあったが、必要な記述情報を抽出できることが確認できた。また、データ抽出の際、適切な匿名化、暗号化などができる仕組みも検討する必要があることがわかった。今後、病院情報システムから、SOAPや退院サマリ、種々の検査報告書など、必要な情報を簡便に抽出できる仕組みとして、基幹システムや部門システムのデータを集約・管理できる統合DBの開発を行う予定である。

電子カルテシステム内の記述情報（SとO）から､自然言語処理および機械学習を用いて、「症状記載」と「その他の記載」の自動判別器を作成した。得られた自動判別器は、正答率80%以上、感度・特異度ともに70～80%の性能を有していた。今後は、MACEに関連するイベントを精査し、そのイベントの判別に必要な教師データの作成や質の高い学習データの作成に必要な医療用語辞書のチューニングを行う。また、SVMだけではなく、ディープラーニングなどいくつかの機械学習手法による比較評価および最適な手法の検討を行う。

G．研究発表

1．論文発表

1）平松 治彦:医療情報システムのデータ利用における課題，Jpn Pharmacol Ther (薬理と治療)，Vol.45 suppl.2，s76-s78，2017

2）平松 治彦:【改正個人情報保護法】医学研究編　国際共同研究など外国にある第三者へのデータ提供について注意すること，医療情報学，37（5），253-5，2017.

3）櫻井理紗、竹村匡正、山口雅和、中井隆史、宍戸稔聡、平松治彦、山本剛、奈良崎大士、上村幸司:ICFを用いた健康情報基盤構築のためのデータ集積手法の検討，第37回医療情報学連合大会論文集，788-789，2017

4）櫻井理紗、竹村匡正、粂直人、岡本和也、黒田知宏：我が国におけるopenEHR/アーキタイプを用いた診療データベースの構築可能性の検証，Mumps，vol.28，15-23，2017

5）山田ひとみ、竹村匡正、桑田成規: 電子カルテの質向上のための診療録監査支援システムの試験的構築，Mumps，vol.28，3-13，2017

6）山田ひとみ、竹村匡正、岡本和也、黒田知宏、桑田成規: インフォームド・コンセント記載を対象とした診療録監査システムの検討，日本診療情報管理学会誌 29(1)，53-61，2017

2．学会発表

1）平松治彦:シンポジウム1「pragmatic clinical trial への誘い」医療情報システムのデータ利用における課題，日本臨床試験学会第8回学術総会

2）櫻井理紗、竹村匡正、山口雅和、松本佳久、本谷崇之、今津貴史、上村幸司、平松治彦、山本剛、奈良崎大士、宍戸稔聡: ICFを用いた個人健康管理システムの構築，第44回日本Mテクノロジー学会大会

3）櫻井理紗、竹村匡正、山口雅和、中井隆史、宍戸稔聡、平松治彦、山本剛、奈良崎大士、上村幸司: ICFを用いた健康情報基盤構築のためのデータ集積手法の検討，第37回医療情報学連合大会

H.知的財産権の出願・登録状況

1．特許取得　　なし
2．実用新案登録　　なし
3．その他　　なし