平成28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・平松　治彦・上村　幸司・竹村　匡正

研究要旨：

リアルワールドにおける電子カルテ記事の分析を行うための下準備として、(1) 自然言語処理のためのフレームワーク、および (2) 病院情報システムからのデータ抽出方法について検討を行った。(1)形態素解析器としては、Juman++の有用性が明らかとなった。今後は、収集したカルテ記載データを用いた検証を行い、その結果を医学用語辞書のメンテナンスへフィードバックしていく必要がある。(2) 病院情報システムからのデータ抽出については、電子カルテ記載情報のうち、SOAP 情報は DWH から抽出できることが確認できたが、報告書やサマリは、部門システムごとに抽出する必要があり、抽出方法については引き続き検討を要することがわかった。また、テキスト情報を失った PDF 形式のデータが多く存在していたため、自然言語処理を行う前処理としての OCR などの検証が今後必要であることがわかった。

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・サマリなどのテキスト情報から、自然言語処理および機械学習を用いて、Major Adverse Cardiac Event（MACE）とよばれる主要有害心血管イベントを予測するモデルを構築することである。

テキスト情報を対象に機械学習をするためには、記載内容を自然言語処理し、適切な用語を抽出する必要がある。しかし、精度の良い用語抽出には、最適な自然言語解析器の選択や医学用語辞書のチューニングをしなくてはならない。また、対象とする所見・報告書・サマリなどのテキスト情報は、様々なシステムで作成・保管され、その形式も多様であるため、本研究で利用できる形式で抽出・収集可能か検証が必要である。

そこで、今回は、実際の電子カルテ記事の分析を行うための前段階として、1. 自然言語処理のためのフレームワークおよび、2. 病院情報システムからのデータ抽出方法について検討を行った。

B. 研究方法

1. 自然言語処理のためのフレームワーク

形態素解析器および医学用語辞書の精度向上のための検討を行った。検討の際には、形態素解析の精度・解析結果の取得のしやすさ・実装のしやすさ・解析速度・解析結果の安定性・医学用語辞書データのメンテナンス性などの観点から評価した。

2. 病院情報システムからのデータ抽出方法

対象となる所見・報告書・サマリなどの情報が、どのシステムにどのような形式で保管されているか確認し、抽出方法を検討した。

（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．自然言語処理のためのフレームワーク

形態素解析器については、申請時には MeCab の利用を想定していたが、本年度に京都大学で開発された Juman++の性能が高いことがわかり、本システムの利用を検討した。その結果、解析器の精度・利便性などの面から十分に利用可能と考えられた。

また、医学用語辞書については、継続的に手作業でメンテナンスを行い、医学分野における形態素、構文情報、シソーラスなどの整備を行った。

2．病院情報システムからのデータ抽出方法

電子カルテに記載されている所見（いわゆる SOAP 情報）は、データウェアハウス（DWH）のデータベースから抽出可能なことが確認できた。

一方、放射線・心エコー・心電図などの各種検査報告書は、各部門のレポートシステムで記載・保管された後、PDF 化されて電子カルテシステムから参照できる形で保管されていた。

同様に、診療情報提供書や退院サマリも専用の文書作成システムにおいて作成され、印刷・押印・スキャンを経て、PDF 形式で保管されていた。

また、手術記録は、作成方法が統一されておらず、Word や Excel などで記載・印刷されたのちスキャンされて PDF 形式で保管されていることがわかった。

このように、記載情報の多くが、テキスト情報を失った PDF 形式で保管されていることが確認できた。

D．考察

1．自然言語処理のためのフレームワーク

JUMAN++は、言語モデルとして Recurrent Neural Network Language Model (RNNLM) を用いることにより、単語の並びの意味的な自然さを考慮することが可能な解析器である。そのため、JUMAN、MeCab に比べ大きく性能が向上している。今後、取得された実際の電子カルテ記載データ、特に検査値などの数値情報が混在するテキストデータでの検証を行い、解析精度などの検証を行う予定である。また、その結果を踏まえて、医学用語辞書のメンテナンスを手作業で継続して行う。

2．病院情報システムからのデータ抽出方法

電子カルテ記事のうち、SOAP 情報は DWH のデータベースから抽出可能なことが確認できたが、その他の情報は、多くがテキスト情報を失った PDF として保管されており、そのままでは自然言語解析を行うことができないことがわかった。

報告書やサマリなどは、各々を作成した部門システムのデータベースから直接抽出する必要があるが、その際には、日常業務に差し障りのないように、システム負荷や時間帯などを考慮した安定的な取得方法を検討しなければならない。

また、手術記録はスキャンされた PDF ファイルしか存在しなかった。したがって、自然言語処理を行うためには、前処理として OCR 等でテキスト情報への変換が必要である。その場合、OCR の精度が問題となるため、今後十分な検証を要する。

E．結論

本年度は、実際の電子カルテ記事の分析を行うための下準備として、(1) 自然言語処理のためのフレームワークおよび (2) 病院情

報システムからのデータ抽出方法について検討を行った。形態素解析器としては、Juman++の有用性を評価することができた。今後は、収集したカルテ記載データを用いた検証を行い、その結果を医学用語辞書のメンテナンスへフィードバックしていく必要がある。電子カルテ記載情報のうち、SOAP 情報は DWH から抽出できることが確認できたが、報告書やサマリは、部門システムごとに抽出する必要があり、抽出方法については引き続き検討を要する。また、テキスト情報を失った PDF 形式のデータが多く存在していたため、自然言語処理を行う前処理としての OCR などの検証も必要であることがわかった。

F. 健康危険情報

なし

G. 研究発表

1. 論文発表

なし

2. 学会発表

なし

H.知的財産権の出願・登録状況（予定を含む）

1. 特許取得

なし

2. 実用新案登録

なし

3. その他

なし