博士(医学) 西本尚樹

学位論文題名

Developing and evaluating an algorithm for the medical concept retrieval in secondary use of radiology knowledge

(医学概念を抽出する基盤技術の開発と評価に関する情報学的研究
－放射線医学知識の二次利用を目指して－)

学位論文内容の要旨

【Background and objectives】 1990 年代から病院情報システムが普及するにつれて、日々様々な形式のデータが保存されるようになった。データによっては、医師によって記述された所見などが自由記述の形式で保存されている。CT レポートや退院時サマリーといった文書は、それ自体が医学概念や専門家の判断を含む豊富な知識資源であり、電子化の歴史も長い。コンピュータによって、これら自由記述で保存された情報を扱うことができれば、診断支援や疾患と習慣の疫学的な調査、医学用語の関係を可視化することによる教育への応用など種々の利点が挙げられる。

しかしながら、自由記述のデータはその量も膨大であり、いまだ解析が困難である。その原因は、(1)用語・語彙の問題、(2)解析ツールの未発達、(3)文法など言語特有の問題に分けられる。(1)については、知識の再利用を促進する活動のひとつとして語彙の標準化が挙げられる。米国国立衛生研究所は、統合医学用語システム(UMLS)と呼ばれる複数の用語集をまとめ、相互に関連づけた用語システムを開発した。(2)と(3)については、臨床医学文書から医学用語を適切に抽出するための専用ツールの開発が望まれている。本研究では、医学用語の意味を抽出する専用ツールが存在しないという問題を解決することに着目した。

本研究の主要な目的は、読影レポートから医学概念を適切に抽出するために、読影レポートに特化した手法論を構築することである。この目的を達成するために、本学位論文では 5 項目の実証実験について記述した。

2 章：用語の意味を扱うオントロジーと呼ばれる学問領域のうち、放射線医学領域に関連する文献調査を行い、含まれる概念を定量的に評価した。医学生物学の概念が多い(36.9%)ことが明らかになった。

3 章：汎用自然言語処理ツールと UMLS を用いて CT レポートの意味分布を可視化した。解剖名(40.6%)、疾患名(20.7%)に関する概念が多いことが明らかになった。

4 章：CT レポートに特化した用語分割のアルゴリズムを開発した。従来法よりも高い精度(81.0%)で医学用語の分割が可能であった。

5 章：用語分割誤りの修正アルゴリズムを構築した。ルールベースの手法により、医学用語分割の精度向上(90.7%)が可能であった。

6 章：意味情報を付与された CT レポートコーパス（文書データ）を作成した。UMLS による意味分布と比較し、一貫性のある結果が得られた。

この抄録では、4 章と 6 章について記述する。

【4 章】

[Materials and Methods] 2005 年 7 月に北海道大学病院で作成された CT レポート 2,000 件から 100 件をランダムに抽出して、アルゴリズムの開発を行った。

従来法は、医学用語の分割に単語の出現確率を用いた手法であり、隠れマルコフモデルと呼ばれる。隠れマルコフモデルは単語に基づく分割であるため、用語集にない「未知語」が文に発生していると適切に区切ることができない。隠れマルコフモデルでこの問題に対応するには、辞書の語彙数を膨大にした状態で維持しなければならず、そのメンテナンスは現実的ではないという問題点が挙げられる。

そこで、文字列の可逆圧縮アルゴリズムである PPM(Prediction by Partial Matching)と呼ばれる手法に注目した。提案する PPM は単語ではなく、1 文字ずつの接続確率に適用可能であるため、医学用語を分割するためにあらかじめ用語集を用意する必要がない。現在までに文字列圧縮アルゴリズムである PPM を医学用語の分割に適用した例はない。
文字の接続確率のみを用いた PPM と、既存の単語に基づく手法によって医学用語分割の性能を比較した。評価指標は、適合率と再現率を用いた。正解となる規準データは、二名の放射線科医によって手動で単語分割を行った結果を採用した。

[Results and discussion] 総文字数 26,036 字から 646 種類の語彙が得られた。適合率は、既存の単語に基づく手法が 51.8%であったのに対して PPM は 81.0%であった($p<0.05$)。再現率は、それぞれ 67.0%、77.6%であった($p<0.05$)。

適合率と再現率の向上によって、経験のある放射線科医による用語分割と PPM による用語分割が近いことが明らかになった。用語分割の精度や用語集のメンテナンスの有無は、後の意味処理に影響を及ぼす。新聞などの一般的と考えられる文書の解析精度が 95%以上を達成していることから、さらなる精度の向上にはルールベースによる処理との組み合わせなどが考えられる。

[Conclusion] CT レポートの解析には、単語に基づくアルゴリズムよりも文字に基づく解析アルゴリズム PPM*の方が有効である。しかし、PPM*の結果は、汎用的なツールで新聞などの文書を解析した精度には及ばなかったため、今後精度向上に向けた工夫が必要である

【6 章】
[Materials and Methods] 同じ CT レポート 2,000 件から 40 件をランダムに抽出し、意味情報として、北米放射線学会が公開している放射線医学分野の標準用語集 RadLex を用いた。尚、RadLex は階層構造を持ち、上位の語彙がより抽象的な用語を表している。意味情報を付与する処理を支援するツールを開発すると共に、CT レポートに意味情報を付与した。付与された意味情報から CT レポート中の意味分布をカウントし、先行研究より UMLS の意味情報を付与した結果と比較した。

[Results and discussion] RadLex の上位語彙として 1,129 語が得られた。頻度の高い語彙から順に、"anatomic entity" (44.8%)、"imaging observation" (26.6%) 、 "imaging observation characteristic" (24.9%)が得られた。"anatomic entity"には"portal vein"や"liver"が、"imaging observation"には疾患名、"imaging observation characteristic"には方向や放射線透過性に関する用語が含まれていた。

解剖名と疾患名について高頻度であることから、UMLS と RadLex で一貫性のある結果が得られたと考えられる。UMLS が英語で 27 万語以上を、RadLex は 12,515 語を収録しているが、RadLex が放射線医学領域に特化しており、CT レポートの用語を抽出する上で適切な用語集であることが明らかになった。6 章では、人手により意味情報を付与した。構築した CT レポートコーパスは、読影レポートから医学概念を自動抽出する際に、ツールの評価指標となる。

[Conclusion] CT レポートの用語を抽出する上で RadLex が適切な用語集であることが明らかになった

【Conclusion】 本研究により、読影レポートから医学概念を適切に抽出するために、読影レポートに特化した手法論を構築することが可能であった。本技術で大量のテキストを自

動処理することにより、以下の例を持って医学に貢献できると考えられる。
(1) 疾病統計など、自由記述の文章から計数可能なデータを抽出する
(2) 喫煙習慣の調査など、ある疾患に付随する症状・習慣の retrospective survey
(3) RECIST 規準に基づく効果判定の計算など、ルーチンワークの軽減

学位論文審査の要旨

主　査　教　授　岸　　玲子
副　査　教　授　福田　　諭
副　査　教　授　白土博樹
副　査　教　授　寺沢浩一
副　査　教　授　玉木長良

学位論文題名

Developing and evaluating an algorithm for the medical concept retrieval in secondary use of radiology knowledge

(医学概念を抽出する基盤技術の開発と評価に関する情報学的研究
－放射線医学知識の二次利用を目指して－)

本研究の目的は、読影レポートから情報を抽出し２次利用を促進するための基盤技術を開発するものである。具体的には、CT レポートから医学用語を抽出するための特化したアルゴリズムを開発することである。本研究の成果は、診断支援、疫学調査の事前情報の提供や予防情報の提供などに応用可能である。

医学用語の適切な分割位置を求めるため、文字と文字の遷移確率から CT レポートの医学用語を分割する手法を提案した。提案する手法は、医学用語の分割に用語集を使用しない。文字列圧縮アルゴリズムを医学用語の分割に用いた研究は過去になく、自由記述の臨床医学文書を定量的に解析するための新しいアプローチである。2005 年 7 月に北海道大学病院で作成された CT 読影レポート 2,000 件から 100 件をランダムに抽出して、アルゴリズムの開発を行った。申請者は文字列の可逆圧縮アルゴリズムである PPM(Prediction by Partial Matching)と呼ばれる手法に注目した。得られた文字間の遷移確率のみを用いて PPM と既存の隠れマルコフモデルによる医学用語分割の性能を比較した。

評価指標は、適合率(precision)と再現率(recall)を用いた。正解となる規準データは、二名の放射線科医によって手動で単語分割を行った結果を採用した。結果として、適合率は従来法が 51.8%であったのに対して PPM は 81.0%であった。再現率は、従来法が 67.0%であったのに対して PPM は 77.6%であり、10%以上の向上が見られた。これらの結果より、読影レポートから知識抽出を行うための基盤技術の開発及び精度を明らかにした。

質疑応答では、白土教授から開発したアルゴリズムについて、着想の経緯に対する質問があった。スライドに研究の全体像を示し、本論文の第３章で研究の開始時に意味解析を行ったことを回答した。第３章で、既存ツールの組み合わせによる意味解析には精度向上の限界があり、CT レポートに特化した全く新しいアルゴリズムの必要性が明らかになったことを説明した。また、白土教授から用語分割の問題をどのように定式化したのか、臨床応用の可能性について質問があった。申請者はスライドに PPM という手法のアルゴリズムを示し、一文字に対して、一つ前の文字との間に医学用語の区切りが入るか入らないかの２つの状態の遷移であると説明した。用語の区切り方の候補は何通りも出ることを示した後、最適な候

補を尤度によって決定していることを述べた。申請者は、臨床応用について研修医が 2 ヶ月間に記述した読影レポートを解析する例を述べ、臨床教育に応用可能であることを述べた。

福田教授からは、CT レポートの部位に偏りはなかったかについて質問を受けた。申請者は CT レポートの部位による分類数を提示し、本研究の課題であることを説明した。また、福田教授より診療科による記述の差異を吸収できるのか、疾患名や TNM 分類を一義的に自動認識することは可能かという質問があった。申請者は、PPM を実装したツールの設計について解説し、参照する文字の遷移確率データを変更すれば、核となる PPM の部分を変更することなく複数科の記録に対応することが可能であることを説明した。また、カルテの記載内容の一部である TNM 分類については、データ量が蓄積して頻度データを更新することで適用であることを説明した。

寺沢教授からは、英語と日本語による言語解析の困難さに違いはあるのかという質問があった。言語学の学会で議論されている見解を引用し、言語による困難性の違いは解析するフェーズ（用語解析、係り受け解析、意味解析）の違いであることを説明した。

岸教授は、公衆衛生学の視点から、疫学的な調査とはどのようなことを指すのか、予防情報とはどういった情報を抽出するのかという質問があった。申請者は、本研究が疫学調査のデータになるわけではなく、疫学調査を行うために疾患と症状、疾患と習慣の相関を可視化することに役立つことを説明した。申請者は追加のスライドを用いて米国で行われている大規模な先行研究である i2b2 challenge を紹介し、習慣の例として喫煙の状態が自由記述文書から抽出される先行研究を示した。

いずれの質問に対しても、臨床応用について具体例を提示し、学位論文を引用して適切に回答を行った。

この論文は、Methods of Information in Medicine で高く評価され、臨床医学文書の自動解析への発展が期待される。審査員一同は、これらの成果を高く評価し、大学院博士課程における研鑽や取得単位なども併せ申請者が博士（医学）の学位を受けるのに充分な資格を有するものと判定した。（1,949 字）