博 士(医 学) 上 杉 正 人

学 位 論 文 題 名

医療用語の意味解析における数理科学的手法の適用

—医学・医療文書から知識体系化を目指して—

学位論文内容の要旨

【背景と目的】

コンピュータの発達により医療分野においてもIT化が普及し、病院や診療所にはさまざまな医療情報システムが導入されている。これらのシステムには診療録、読影所見や退院時要約など自由記述の医療情報が保存管理されている。これらの情報の二次利用を考えたとき、医療用語が使用者によって異なったりあるいは表記に揺れが生じるため、そのままではコンピュータが処理できない問題が存在する。用語の不統一と表記の揺れは医療情報の二次利用にとって大きな問題である。この問題を大量に保存された医療文書から用語の出現頻度を計測し、数理科学的手法を用いて用語を抽出する方法の研究を行った。また、自由記述の医療文書を分類し体系化することは医療・医学知識の構築のために必須になると考える。しかし、人間が分類するには高度な専門的知識と膨大な労力が必要になる。更に、分類結果は作業する人間の背景知識と経験により恣意的になる可能性がある。そこで本研究ではコンピュータによる自由記述文書の自動分類を数理科学的手法により再現性良く、自動的に分類する手法について研究を行った。また、本研究は自由記述の医療文書から知識を抽出して体系化するための基礎的研究も含んでいる。

【対象と方法】

解析対象とした医療文書は医学中央雑誌刊行会の症例報告の抄録から無作為に抽出した9800例(321万文字)である。対象文書中の文字列の出現確率であるN-gramは1文字から4文字までの組み合わせを求めた。これらのN-gramを組み合わせ6種類の相互情報量と各変化量を計算した。そして6種類の相互情報量に重み係数をかけた合計値がある閾値以下であり、また6種類の相互情報量の変化も重み係数をかけた合計値がある閾値以下であるとき用語の分割境界とした。最適な12個の重み係数と2つの閾値を求めるため、遺伝的アルゴリズムを用いて学習を行った。学習用医療文書中の736語の医療用語が最も効率的に分割できる最適な重み係数と閾値を求めた。

自由記述の医療文書の分類の基礎として、粒度を一つ下げ医療用語の自動分類を行った。対象とした医療用語は雑誌「循環器」の抄録タイトルから抽出した循環器病名61語と「日本消化器科外科雑誌」抄録タイトルから抽出した消化器病名82語の合計143語である。143語は一文字一文字に分解され、共起行列を作成した。つまり行に医療用語とり列に文字をとり、用語にある文字が出現するとそのカラムに1が加えられる。この共起行列から潜在的意味解析(Latent Semantic Analysis)を用いてセマンティックスペースに展開した。各用語の行ベクトルの適合度(Cosine)を求めることにより用語の意味的な近さを計測した。適合度は+1から-1までの範囲をとり、+1に近いほど意味的関係が近く、-1に近いほど意味的関係が遠いことを示す。対象とした用語間の適合度が1.0〜0.8、0.8〜0.6と0.6〜0.4の3つのレベルで用語の関係を図示した。

【結 果】

医療用語の分割において、遺伝的アルゴリズムを用いて 20,000 世代の計算で最適解に収束し、12 の重み係数と 2 つの閾値を求めた。学習に使用した 736 語のうち 600 語(81.5%)で用語の分割境界を正しく探索した。この内、371 語(50.4%)は想定した 1 語で正しく分割された。また 194 語(26.4%)は“造影|不良”のように 2 語以上に分割されたが、分割されたそれぞれの語は意味を持っていた。しかし 35 語(4.8%)は分割されすぎ“乏|血|性”のように意味不明の分割が存在し、カタカナ表記やアルファベット表記に現れる傾向があった。

用語の自動分類において、循環器病名は適合度が 0.4 以上のとき、56 語(91.8%)が 14 のグループに分類された。適合度が 0.4 未満の循環器病名は 4 語存在した。また、消化器病名は適合度が 0.4 以上で 71 語(86.6%)存在し、16 のグループに分類された。適合度が 0.4 未満の消化器病名は 5 語存在した。また、6 語の消化器病名は循環器病名に意味的に近いことを示した。

【考 察】

医療用語の分割において、最適な 12 個の重み係数と 2 個の閾値が得られた。これらの最適解を用いて 1 語または 2 語以上に分割された 465 語(76.8%)は有効な分割であった。しかし、薬剤名などのカタカナやアルファベット表記は過剰に分割される傾向があった。また、正しく分割できなかった 136 語(18.5%)のうち 79 語(10.7%)は用語に格助詞あるいは記号が付いていた。これらの過剰な分割と余計な語が付加される問題を解決するために、前処理と後処理を加えることにより分割の精度が向上できると考えられる。前処理にはカタカナあるいはアルファベット表示の文字列は一語として処理する。後処理には、分割された用語に格助詞または記号が付くとき、付加された文字を取り除く処理を行う。この 2 つの処理を加えることにより分割精度を 92.0%に向上することができる。

用語の分類において、循環器病名と消化器病名全体において 127 語(88.8%)が正しく分類することができた。これは病名が複合語で構成されているため、医療用語に含まれる文字がそれぞれの領域で類似していたと考えられる。このことは医療用語の表記の揺れなどによる曖昧な用語に対しても分類することができる。例えば、“冠状動脈疾患”、“冠動脈疾患”や“冠状動脈性心疾患”などの表記の揺れは容易に判断できる。一方、正しく分類できなかった 16 語については、この内 9 語が他に共起する用語が無く、また 7 語に関しては他の病名群と関連を示していた。他の病名と関連を示した用語では、消化器病名群の“門脈血栓症”は循環器病名群の“冠動脈血栓”に関連性を示した。お互いにの用語に含まれる“脈血栓”の共起が原因と思われる。今回の消化器病名群に“門脈”を含む用語が無かったためこのような現象が起きたと考えられる。このように用語の集団の偏りにより、分類の結果が影響を受けることが分かった。今後、用語分類に関しては用語集団の偏りに影響を受けないように集団の構成を検討する必要がある。つまり、大規模な用語群を用意し処理する必要がある。

【結 論】

医療の IT 化により大量に保存された自由記述のテキスト情報から数理科学的手法を用いて用語を分解抽出し、意味的関係を求めることができた。この研究は数理科学的手法を用いているため、人間の労力を必要としないのはもちろんだが、結果について再現性がある。今後、さらに大規模な自由記述の医療文書に対して数理科学的な解析手法を応用することにより医療用語抽出精度を向上させ医療用語辞書を自動的に構築できる可能性がある。また日々進歩する医学医療の変化に対応した解析技術であると考える。更に、医学情報を効率的に集めて意味的情報を整理し体系化することにより、医療の中での判断支援を支える真の IT 化が実現できると考える。そのため本研究は数理科学的手法が有用であることを示した。

学位論文審査の要旨

主　査　教　授　玉　木　長　良
副　査　教　授　白　土　博　樹
副　査　教　授　寺　沢　浩　一
副　査　准教授　遠　藤　　晃

学位論文題名

医療用語の意味解析における数理科学的手法の適用

—医学・医療文書から知識体系化を目指して—

医療分野におけるIT化の普及により医療情報システムの導入が増え、診療録、読影所見や退院時要約など自由記述の情報が大量に医療情報システムのデータベースに保存管理されている。また、インターネットの発達により医学・医療の有用な情報も多く存在している。これらの医療情報を単にキーワードだけによる表層的な検索ではなく、用語の関係性を数理科学的な手法による意味解析を視野に入れた医療知識ベースを構築して医療支援システムの開発への挑戦がこの論文の研究の背景である。

本論文は5章から構成されている。第1章の序論では、医学用語と概念の関係について述べ、医療用語が臨床の現場で用いられるとき、分野により用語の持つ意味的な背景概念が異なり、また時代とともに変化することを示している。こうした問題を人や辞書を用いた解析システムで捉えることは難しく、数理科学的モデルを用いて解決することを提案している。第2章では医療用語の特長について研究している。本研究では新聞社説と症例報告文書との比較において医療用語は長い用語が多く、また基本となる用語が繰り返し出現する性質、つまり複合語であることを明らかにしている。第3章では医療用語の分割において数理科学的手法を用いて研究している。本研究では、医療用語は利用者による表記方法が違うことや、表記の揺れの存在を前提に、従来の医療辞書を用いず、医療文書の文字間の特性により医療用語を抽出する方法を提案した。そのアルゴリズムは大量に保存された医療文書から用語の出現頻度であるN-gramを計算し、4種類のN-gram(Uni-gram, Bi-gram, Tri-gram, Quadri-gram) から6種類の相互情報量とその変化量を用語分割の指標とし、最適な用語分割のパラメータ（指標の重み係数と閾値）を得るために遺伝的アルゴリズムを用い、その結果有用な医療用語の分割が76.8%の精度で達成でき、またアルファベット表記やカタカナ表記の文字列を1語として取り出す前処理や、正しく分割できなった格助詞や記号を取り除く後処理を加えることによって91.9%にまで精度を向上できることを示した。次に医療知識ベースの構築のためには情報を分類・整理することが重要であることが述べられ、この作業には専門家の協力が必要であるが、膨大な労力が伴うだけではなく、分類整理に伴う判断が専門家の経験や専門分野により揺らぎ一貫性に欠けることが指摘されている。この問題を第4章の医療用語の意味解析において数理科学的モデルを用いて解決で

きることを研究している。この数理科学的モデルは潜在的意味解析と呼ばれ、本研究では循環器病名 61 語と消化器病名 82 語を対象に自動分類を試みている。その結果、潜在的意味解析は 91.8%の循環器病名と 86.6%の消化器病名を正しくそれぞれの領域に自動分類し、さらに用語の意味関係から循環器病名は 14 グループに分類し、消化器病名は 16 のグループに分類できた。本研究は数理科学的手法が医療用語の自動分類に有効であることを示した。第 5 章では結語としてこの研究の意義などに言及し、医療文書に数理科学的手法を用いて医療知識ベースを構築して、医療支援、医学教育や疫学研究に応用できることを示した。

口頭発表に際し、副査の白土教授から、画像解析と言語解析の関連性についての質問がなされた。また、臨床で安全にシステム使うためには真の値が必要であり、その評価についての質問があった。副査の遠藤准教授から市販のテキストマイニングとの比較についての質問があった。さらに、N-gram を 4 文字以上に増やした場合の効果の改善についても質問があった。副査の寺沢教授から本手法の再現性について、何と比較したのかについて質問があった。最後に、主査の玉木教授から医療で使っている用語は、他の分野の言語とどういう違いがあるのかに関する質問がなされた。さらに、英語と日本語での医療用語の違いについての質問があった。いずれの質問に対しても、申請者は研究結果および文献的知識により、概ね適切に回答した。

この論文は医療用語の特徴を調べ、用語の出現の確率と相互情報量を用いて医学・医療文書から適切に用語を抽出し、また潜在的意味解析手法を用いて用語の意味関係から分類を行なっている。この研究は医療情報を体系つけられた医療知識ベースの基礎となり、医療支援、医学教育や疫学研究への応用の可能性を示唆したということで意義のあるものと評価され、審査員一同は、これらの成果を高く評価し、大学院過程における研鑽や取得単位などもあわせ申請者が博士（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。