[別紙―1]

論文の内容の要旨

論文題目　日本語疾患表現の構造解析とその ICD コーディングへの応用に関する研究

指導教員　大江和彦 教授

東京大学大学院医学系研究科

平成 18 年 4 月進学

医学博士課程

社会医学専攻

氏名　山田恵美子

序文

コンピュータが広く普及した現在，多くの情報が電子的に蓄積され，日々その量を増している．医学医療分野もその例外ではなく，病院情報システムの導入やレセプトオンライン化など政府主導による電子化が進みつつある．情報が電子化された時の利点の一つとして，コンピュータで自動処理することが可能となることが挙げられる．その中でも自然言語で表現されたデータを対象とする自動処理の試みは，主に自然言語処理分野において行われており，その成果として検索エンジンや機械翻訳などのアプリケーションが実社会においても広く認識・利用されている．このような技術は，時間的な制約から人手では現実的に不可能であったことを実現したり，ある作業で必要となる知識の不足を補ったりするものであり，これまでにない価値を生み出している．

自然言語処理での基礎技術である形態素解析（テキストを最小単位に分割して品詞を付与），係り受け解析（文節間の係り受け関係の同定）は広く研究されており，既にそれぞれ 99%，90%程度の精度を達成している．ここから更に高度な処理を行うためには語が示す概念についての知識が必要不可欠である．そのためのリソースとしてシソーラスやオントロジが構築されている．一般的にこのようなリソースは構築の過程で人手を介するので内容の妥当性が保証されるものであるが，一方で膨大な量の語・概念を扱うために，その構築や維持管理には多大な労力が必要となるという問題がある．語の数を多くする要因の一つとして，自然言語は非常に柔軟な表現力を持っており，特に短い語の組み合わせ（複合語）として表現される専門用語は無数に生産されうるということが挙げられる．

そこで，既知の語が組み合わさって構成される複合語に関しては，人手で意味を記述するのではなく，構成情報から意味を推測する，あるいは構成情報を疑似的な意味として扱いたいという要求が生まれる．複合語の構成情報を知るためには，(1)複合語を単位語に分割し，(2)単位語同士の関係（主に修飾・被修飾関係）を同定する，という 2 点が必要である．これは上述の形態素解析および係り受け解析の結果得られる情報とほぼ同等のものと考えられるが，自然言語処理学における解析は文を対象としたものであり，複合語を対象としたものではない．

方法

本研究では，まず医学用語（特に疾患名）の内部構造の表現方法を提案した．次に提案した内部構造情報の医療における有用性を評価するため，これを用いた新しい自動 ICD コーディング手法を提案しその評価を行った．さらに，実際に内部構造情報を利用するために必要となる，文字列で与えられた用語を自動解析して内部構造表現を得る内部構造解析器を作成し，その精度評価を行った．

1. 内部構造表現の提案

本研究において，用語の内部構造とは，語を構成する単位語とその結びつきの順序を示す．提案する内部構造表現は，単位語対の関係集合である係り受け構造としての表現である．単位語の明確な定義を決めるのは非常に難しい問題であるため，本研究では文字を単位語とすることにした．さらに，「心臓＋疾患＝心疾患」，あるいは「角膜＋結膜＝角結膜」や「大腿骨＋骨折＝大腿骨折」のように複数の語が結合する際に文字が抜け落ちる現象（省略/縮退現象）を表現できるよう表現方法を拡張した．具体的には係り受け関係の種類として通常の係り受け関係に加え文字を生成する係り受け関係を追加した．内部構造の決定方法は，語の 2 分割と分割された 2 語間の係り受け関係の決定を再帰的に行う方法を提案した．分割位置は前方・後方探索により決定することとし，係り受け関係における修飾語・被修飾語の決定方法は，合成語との間に上位下位関係または部分全体関係が成立する構成要素を被修飾語とすることとした．

2. 自動 ICD コーディング手法への応用

次に，提案した内部構造表現の有用性を評価するため，自動 ICD コーディングへの適用方法を提案した．ICD コーディングは疾患名に ICD コードを付与する作業であり，そのための資料として ICD コードが既知である疾患名リストである標準病名マスターが公開されている．しかし入力となる疾患名は表記ゆれを伴った自由入力であり，標準病名マスターの中に概念としては収載されていても，文字列としては収載されていない場合も多い．表記ゆれには(1)翻字，(2)同義語，(3)修飾語の順序や有無の 3 つのパターンが挙げられる．提案した内部構造表現は(3)を解消するのに有用であると考えられ，本研究ではこれを対象とした自動 ICD コーディング手法を提案した．提案手法は先行研究の手法を拡張したものであり，以下の 2 段階から構成される．Step1）まず文字列で入力された疾患名を内部構造表現に変換する．この時，表記揺れパターン 3 のうち「修飾語の順序」によるものが解消され，例えば「急性 A 型肝炎」と「A 型急性肝炎」は同一の内部構造表現となる．

Step2) Step1 で得られた内部構造表現に対してコーディングルールを適用する．ルールは「内部構造表現→ICD コード」の形で作成し，左辺で用いる内部構造は右辺の ICD コードが持つ必要十分な情報のみから成るものである．提案手法の評価実験として，先行研究の手法でコーディングできない疾患名のうち 100 例を対象として提案手法による自動コーディングを行った．

3. 内部構造解析

この ICD コーディング方法を自動で行うためには，文字列として入力される疾患名を解析し内部構造表現を自動生成する仕組みが必要である．提案した内部構造は一般に使用される文の係り受け構造と同じ枠組みによる表現であり，従って自然言語処理学分野で研究されている係り受け解析技術を利用して自動解析することが可能である．自然言語処理分野では係り受け解析の研究が古くから行われており，様々なアルゴリズムが提唱されている．本研究では決定的アルゴリズム Shift-Reduce モデルによる係り受け解析の実装である係り受け解析器 MaltParser を用いた．また，内部構造を適切に扱うためには縮退や省略を復元する必要がある．復元のためには，(1)縮退・省略の起きた場所，(2)抜け落ちた文字の二点について特定しなければならない．(1)は提案した内部構造表現の中に明示的に表現されており，係り受け解析の対象範囲に入る．(2)については本研究では扱わないが，予測変換等の手法を用いることで解決可能な問題であると考えられる．MaltParser による内部構造解析の精度評価として，標準病名マスターから 696 語について人手で内部構造を記述し，これを学習・テストデータとして 5 分割交差検定を行った．

結果

自動 ICD コーディング実験の結果，提案手法では 100 例中 8 例がコーディング可能であった．このうち内部構造情報が寄与した例は 3 例であり，解剖部位の部分全体関係を内部構造情報から判断可能であったことがその理由であった．

MaltParser による内部構造解析の精度評価実験の結果，文字対の係り受け関係に対して 95.4%，語に対して 83.7%の精度で解析が可能であった．

考察

提案した内部構造表現によって，従来扱われてこなかった省略・縮退現象を明示的に表現することが可能となった．また上位下位関係・部分全体関係を根拠とした内部構造を，上位下位・部分全体関係に関する外部知識を用いずに自動解析可能であるということから，従来ならば外部知識を別途用意しなければ解けなかった問題を，外部知識ではなく内部構造表現を用いることによって解決することができるようになった．本研究で提案した内部構造表現は ICD コーディングのみならず，様々な場面でテキスト処理の新たな基盤となる可能性を持つものであり，今後これを洗練していくことが必要である．具体的な課題として，本研究では扱いの難しい複雑な語を視野に入れた表現方法の拡張，省略・縮退によって抜け落ちた文字の復元，内部構造解析精度の向上のためのアルゴリズムや学習素性の改良が必要である．

結論

本研究では医学用語，特に疾患名の内部構造の表現法として文字単位のラベル付き係り受け表現およびその人手による解析法を提案した．この表現法は従来の形態素による表現では無視されてきた省略・縮退現象を扱うことができるという点で新規性があるものである．また，内部構造を考慮することで既存の自動 ICD コーディング手法では解けなかった問題を解くことが可能になる例を示した．医学用語の内部構造は ICD コーディングのみならず，柔軟な形態素解析，入力支援など，さまざまな場での活用が可能である．

更に，疾患名に対して内部構造を自動付与する解析器を作成した．実験の結果，解析精度として，文字対では 95.4%，語では 83.7%という高い性能を達成した．この結果により，文字列として入力された疾患名の内部構造情報を活用したアプリケーションが実現可能であることを示した．