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

山田恵美子

目次

要旨

電子的に保存された大量のテキストデータを十分に活用するにはコンピュータによる自動処理が欠かせない．高度な処理には医学用語の意味情報が必要不可欠であるが，存在しうる全ての文字列表現の意味を予め記述するのは非常に労力を要するものである．

本研究では(1)医学用語（特に疾患名）の内部構造の表現方法と(2)内部構造情報を用いた自動 ICD コーディング方法を提案した．(2)について，実験の結果，既存手法では解けなかった問題の一部が解けた．

更に，疾患名に対して内部構造表現を自動付与する解析器を作成した．評価実験の結果，83.7%の精度で語を解析可能であり，提案した内部構造表現に実用性があることが示された．

1. 序文

医療は医師や患者の他，コメディカル，保険者等様々な立場の人間や組織が関係する社会システムであり，臨床情報や保険情報など異なる種類の情報が大量に生成・保存・伝達される場である．このような状況において煩雑な作業の支援や代替を実現する情報処理技術が有用であることは論を待たない． 本研究ではその基礎技術の一つとなる，疾患名の文字列内部構造の表現と自動解析を提案する．

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いる．自然言語とは，コンピュータ処理を前提に文法が定義された言語である機械言語に対して，人間が使用するために自然発生的に出現した言語であり，明示的な文法定義が存在しない．自然言語をコンピュータで扱うのは大きな挑戦であるが，近年の研究成果はめざましく，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．

日本語における自然言語処理の基本的な流れは，まずテキストを意味的最小単位である形態素に分割して品詞を付与し（形態素解析），形態素をまとめ上げて文節を作り，文節の間に係り受け関係を付与する（係り受け解析）．解析の際には外部知識として品詞情報の付与された用語辞書が必要である．代表的な形態素解析器として JUMAN[1]，ChaSen[2,3]があり，処理精度は 99%程度（F 値），係り受け解析器としては KNP[4,5]，CaboCha[6-8]があり精度は 90%程度（係り受け正解率）と，高精度での解析が可能となっている．

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挙げられる．シソーラスとは語を広義・狭義の関係などで分類した階層構造を持つ辞書であり，オントロジとは厳密に定義された概念および概念間の関係を記述したものである．一般分野では WordNet[9]，医学分野では MeSH[10]，SNOMED-CT[11]，FMA[12]，openGALEN[13]，更に複数のリソースを統合するメタシソーラスとして UMLS[14]が存在している．また我が国においても医学オントロジが構築されつつある．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[15]．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実質的に無数といってもよいほど多様に生成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，複合語の構造を規定している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．複合語の構造を規定する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，既存の自然言語処理学における解析は文を対象としたものが主であり，本研究のように複合語を対象としたものは少ない．

本研究では，複合語としての医学専門用語，特に疾患名に対し，その内部で構成要素となる語の集合が持つ構造（語の間の関係）を係り受け構造として表現し，更に係り受け解析の手法を応用して文字列からその内部構造を導出する自動解析を試みる．

複合語の内部に着目した研究はこれまでにも多く行われており，漢字熟語内の品詞列・係り受けの調査[16]，医学用語を対象とした用語構造解析[17]，意味クラスの共起情報を用いた構造解析[18]，相互情報量を用いた構造解析[19]などが挙げられる．しかしいずれも語を文字の重複なく分割しているのみで，後に述べる省略や縮退を扱っていない．富樫ら[20]は医学用語内部の係り受け構造解析を試みており，形態素レベルでの縮退現象を扱っているが，本研究で扱う文字単位の縮退現象（例：角膜＋結膜→角結膜）は対象としていない．竹内ら[21]の提案する LCS（Lexical Conceptual Structure：語彙概念構造） は同じく複合語の解析を試みているが，扱う対象がサ変動詞を含んだ語であり，医学用語に対しては不十分である．

2. 目的

図 1 に研究のオーバービューを示す．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とその人手による定義方法を提案する(3.1 節)．これは用語内の語の関係を表現するものである．次に提案した内部構造情報の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う(3.2 節)．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の精度評価を行う(3.3 節)．

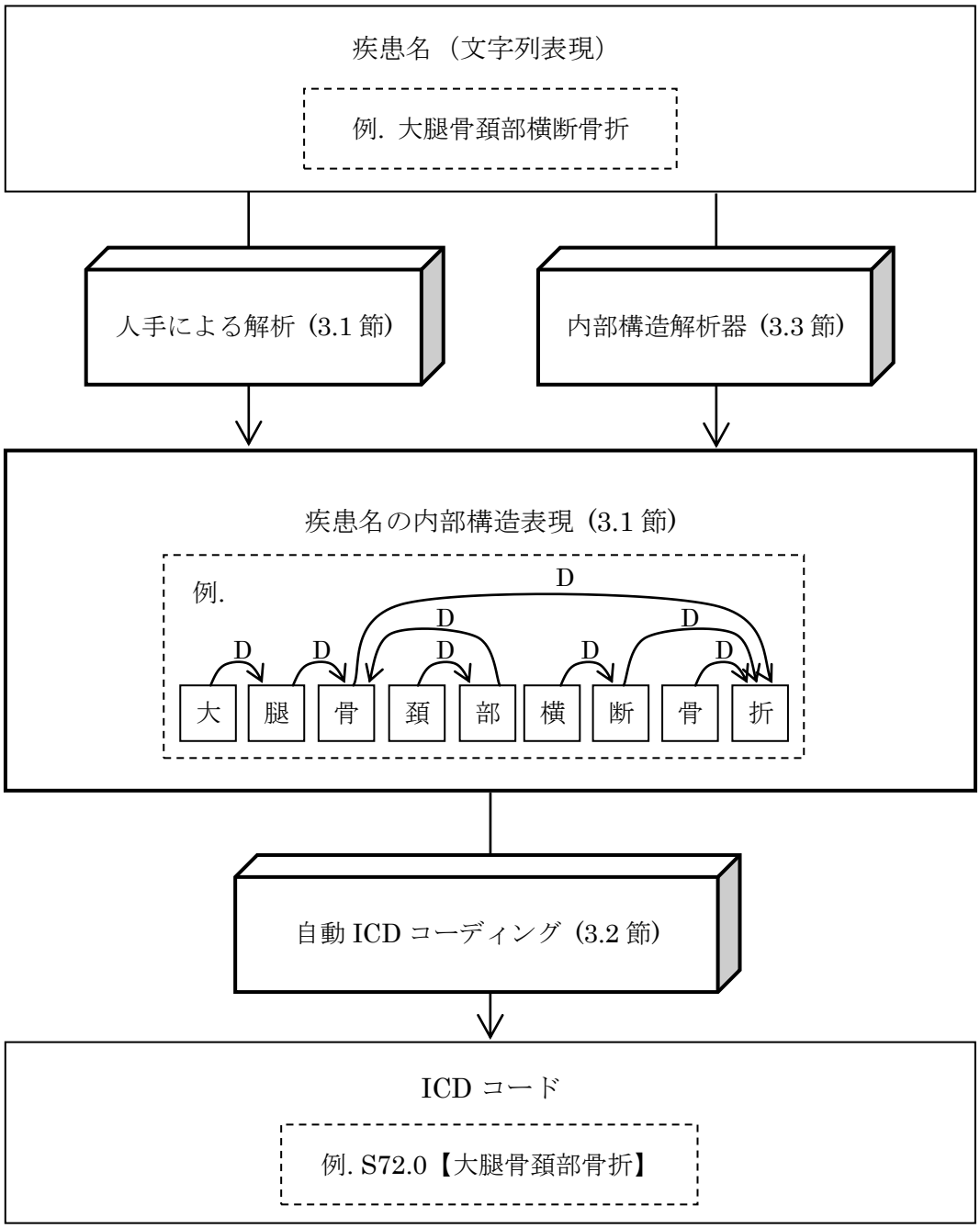


図 1　研究のオーバービュー

3. 方法

本章では，まず本研究の根幹となる医学用語の内部構造表現法とその人手による解析方法を提案する．次にこの内部構造表現法の医療における応用性を評価するため，内部構造情報が事前に得られているという前提のもと，内部構造表現法を使用した自動 ICD コーディング手法を提案し，評価する． 最後に，内部構造表現を文字列表現から自動で生成する手法の実現性を検討するため，機械学習を用いた係り受け解析器の実装の一つである MaltParser による自動生成の精度評価を行う．

3.1. 内部構造表現の提案

3.1.1. 内部構造の基本表現

本研究において疾患名の内部構造とは，疾患名を構成する短い語とその結合順序を意味する．疾患名を構成する短い語とは，それ以上分解できない，すなわち内部構造を持たない文字列である．以下「単位語」と呼ぶ.

内部構造を決定するためには，まず単位語の集合を与えることが必要である．単位語は上記の定義から，意味を有する最も短い文字列，すなわち分解すると意味を失うものであるが，単位語の集合を決めるのは非常に難しい問題であることが指摘されている[22]．例えば「大腿骨」を「大腿，骨」と2つの単位語の列として表現するのが自然である一方で，同じ骨の一種である「肋骨」を「肋，骨」と分割するのは「肋」が独立した語として存在

しないために不自然と感じられる．また「大腿骨」は特定の骨を指し示す言葉であるから分割せず1つの単位語とするべきである，という主張もあるだろう．

このように何を正解とするべきかは利用する人や場面に依存し，唯一の定義は与えられない．本研究ではこの問題を避け，文字を単位語とした．単位語は意味と紐付けられたものであると述べたが，直観的には文字は「何らかの意味を示す単位語」としてはふさわしいものではない．なぜならば独立して存在する一般的な「語」は必ずしも1文字から成るものではなく，また本研究で対象としている日本語では表意文字である漢字だけではなく表音文字である平仮名・片仮名も使用されるためである．したがって本研究における単位語は，「意味的な構成要素」としての単位語ではなく，「文字列の構成要素」としての単位語であり，前者を更に分割したものである．そのため，「意味的な構成要素」としての単位語やそのような単位語の合成によって作られる語は本研究で提案する内部構造には現れないが，「文字列の構成要素」の一部をまとめることで得ることができる．

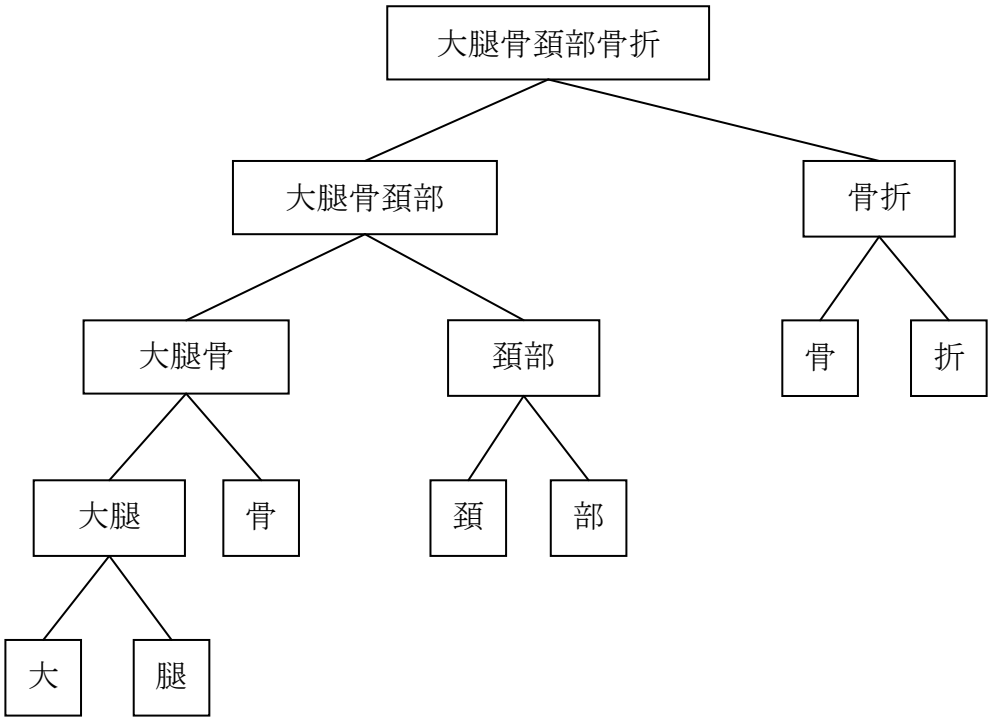


図 2　大腿骨頚部骨折の内部構造（構文構造）.
内部構造に従って単位語（文字）をまとめることで，例えば「大腿」「大腿骨」と正しい語が得られる．語ではない「大頚」は得られない.

図 2 に「大腿骨頚部骨折」の内部構造を示す．これは自然言語処理分野で文に対して適用される構文構造とほぼ同等の表現であり, 語の分割を再帰的に繰り返すことで得られる．本論文では便宜的にこの表現を「構文構造」と呼ぶ．文字列表現から構文構造を導出するためには 1 回の分割において分割位置を 1 か所決定しなければならないが，本研究では以下のルールに従って分割位置を決定することとした.

語の分割位置の決定法

前方・後方探索によりそれぞれの分割候補位置を探す．分割候補位置とは，そこで分割した時に得られる 2 文字列が，分割前の語における意味を保っているものである．得られた 2 つの分割候補位置のうち，Head（後述）が長くなる分割候補位置を正しい分割位置とする．

共に語となる分割が存在しない場合は，末尾 1 文字とそれ以外に分割する．

上記のルール中にある，ある位置が分割候補位置であるか否かの判断は人間が行う．このことに対する反論として，何を単位語とするべきかという議論を避けるために文字を単位語とした内部構造の提案であるにも関わらず，その定義手順に曖昧さを含み得る人間の判断を必要とするのでは，提案する内部構造からは曖昧さを排除できないのではないかというものが考えられる．しかし，単位語の議論が「ある語を単位語列に分割する時，最も妥当な分割を 1 つ選択する」というタスクを対象としているのに対し，上記ルールでの人間による判断は「ある文字列が語であるか否かを判断する」というタスクであり，このタスクは当該領域の知識を持っていれば判断を行う者の間で判断が不一致となる可能性は低いと考えられる．

例えば「大腿骨骨折」の場合，分割位置は以下の 4 か所が考えられる．語として認定されるものを下線で示した．

A) 「大 | 腿骨骨折」

B) 「<u>大腿</u> | 骨骨折」

C) 「大腿骨 | 骨折」

D) 「大腿骨骨 | 折」

前方探索は A から昇順に BCD と，後方探索は D から降順に CBA とチェックしていき，分割後の 2 文字列が共に語となるような分割位置を選択する．この場合，前方・後方探索共に C が選ばれ，C が正しい分割位置となる．

また,「大腿骨頚部骨折」の場合，分割位置は以下の 6 か所が考えられる．語として認定されるものを下線で示した．

A) 「大 | 腿骨頚部骨折」

B) 「大腿 | 骨頚部骨折」

C) 「大腿骨 | 頚部骨折」

D) 「大腿骨頚 | 部骨折」

E) 「大腿骨頚部 | 骨折」

F) 「大腿骨頚部骨 | 折」

C は分割後の 2 文字列が共に語となっているが,「頚部骨折」は分割前の語「大腿骨頚部骨折」における意味とは異なるため，この分割は除外する．その結果，前方・後方探索共に E が選ばれ，E が正しい分割位置となる．

同様にして「甲状腺機能亢進症」に対しては，前方探索で「甲状腺 | 機能亢進症」，後方一致で「甲状腺機能亢進 | 症」の分割位置が選ばれる．後述する基準によりそれぞれの Head は「機能亢進症」「症」となり，Head がより長い前者，すなわち「甲状腺 | 機能亢進症」を正しい分割位置とする．

また，分割後の 2 文字列が共に語となる分割が存在しない場合の例として，例えば「びらん」は「びら | ん」のように分割する．

以上の分割を再帰的に繰り返すことで図 2 のような構文構造が得られる．各分割では必ず 2 つの文字列を生成するので，2 分木法と捉える事が可能である．

次にコンピュータ上での扱いやすさを考え，構文構造を係り受け構造へと変換する．

係り受け構造とは，2 つの単位語の間に成り立つ係り受け関係の集合である．係り受け関係は並列や同格のような場合を除いて非対称であり，係り受け関係にある 2 語は修飾語（Modifier）と被修飾語（Head）の関係にあることが多い．これは構文構造が持たない情報であるので，係り受け構造に変換する際に新たに付加する必要がある．係り受け関係を扱う自然言語処理学分野においては文節の係り受け関係の Head を「複数の構成要素が結びついて出来る語や句の品詞を決定する構成要素」としているが，本研究で扱う疾患名は名詞であり，構成要素も全て名詞であるため，この定義を使うことは出来ない．そこで本研究では以下のように Head と Modifier を定義することとした．

Head と Modifier の決定法

分割前の語（AB）と分割して得られる構成要素（A または B）を比較する.

基準1)　is-a 関係

意味の上位・下位関係（広義・狭義関係）である is-a 関係が成立する構成要素を Head とする

if AB is a A then <Head,Modifier>=<A,B>

if AB is a B then <Head,Modifier>=<B,A>

基準2)　part-of 関係

基準 1 で Head と Modifier が決まらなかった場合，構造の部分全体関係である part-of 関係が成立する構成要素を Head とする

if AB is part of A then <Head,Modifier>=<A,B>

if AB is part of B then <Head,Modifier>=<B,A>

基準3)　is-a / part-of 関係以外

基準 1 と 2 で Head と Modifier が決まらなかった場合, 後ろの構成要素を Head とする

<Head,Modifier>=<A,B>

語の分割位置の決定法と同様，上記ルールにおいても各関係の存在の有無の判断は人間が行う.

例えば,「大腿骨 | 骨折」の場合，まず is-a 関係として以下の 2 つが成立するかどうかを考える.

A) 大腿骨骨折 is-a 骨折

B) 大腿骨骨折 is-a 大腿骨

明らかに A は成立し B は成立しない．すなわち，分割前の語「大腿骨骨折」と is-a 関係にあるのは「骨折」であるので,「骨折」を Head,「大腿骨」を Modifier とする.

別の例として「上腕骨 | 遠位端」の場合，同様に以下の is-a 関係が成立するかどうかを考える.

A) 上腕骨遠位端 is-a 上腕骨

B) 上腕骨遠位端 is-a 遠位端

まず，明らかに A は成立しない．B については,「『遠位端』という抽象概念を『上腕骨』という部位で特化しているから is-a 関係が成立する」と捉える考え方もあるが，本研究では,「『遠位端』は『上腕骨』という文脈なしには存在できない，役割を表す概念（ロール概念）であるから is-a 関係は成立しない」と捉える方針をとった.

次に基準 2 の part-of 関係を考える.

C) 上腕骨遠位端 part-of 上腕骨

D) 上腕骨遠位端 part-of 遠位端

明らかに C は成立し D は成立しない．すなわち,「上腕骨遠位端」と part-of 関係にあるのは「上腕骨」であるので,「上腕骨」を Head,「遠位端」を Modifier とする.

このようにして付与された，語が分割された出来た 2 語の間の関係を，本研究において「係り受け関係」と呼ぶ．構文構造に係り受け関係を付与した内部構造を図 3 に示す．

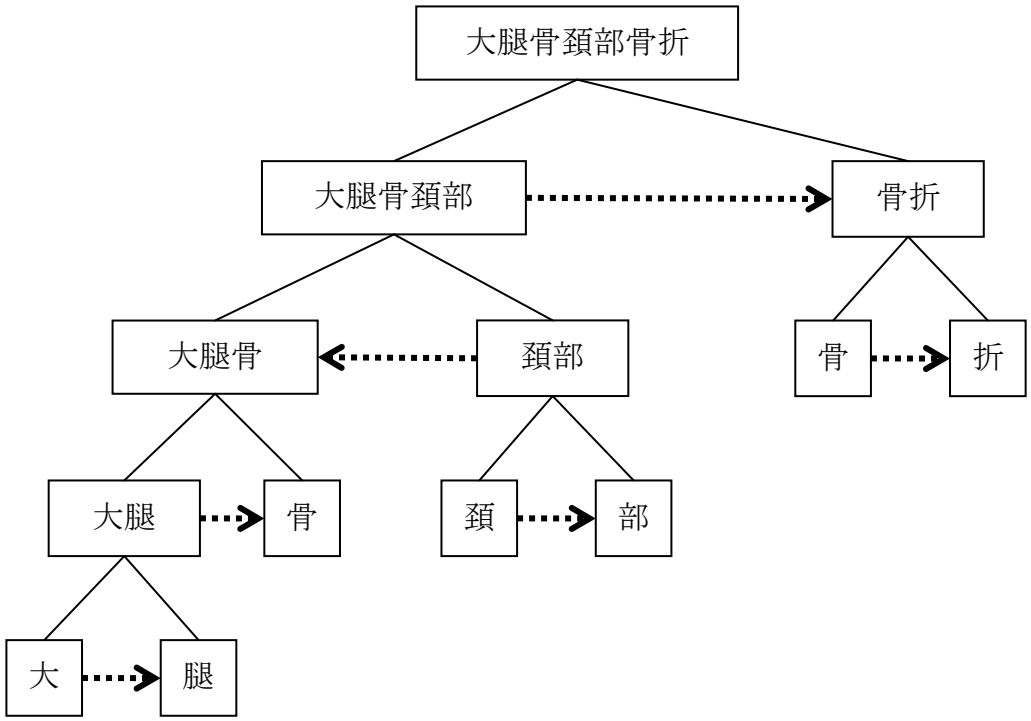


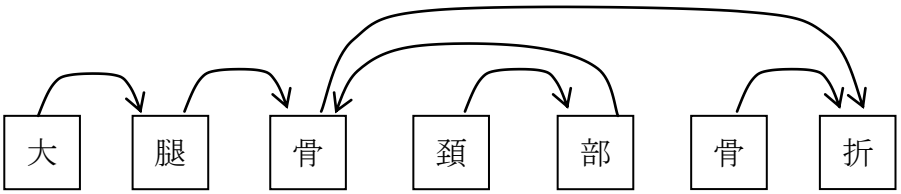
図 3　係り受け情報を付与した内部構造
点線矢印で係り受け関係を示す. 矢印の出る語が Modifier, 入る語が Head である.
例えば係り受け関係「大→腿」において,「大」が Modifier,「腿」が Head である.

次に，これを係り受け構造，すなわち単位語である文字間の係り受け関係の集合として表現する．変換方法は以下の通りである．

係り受け情報を付与した構文構造から係り受け構造への変換法

係り受け関係にある 2 語が共に単位語である場合について，この関係を係り受け構造に含め, 2 語が合成して出来る語をこの係り受け関係における Head と置き換える．これを，係り受け関係に現れる語が全て単位語になるまで繰り返す.

例えば図 3 に対しては，まず単位語間の係り受け関係である「大→腿」「頚→部」「骨→折」を係り受け構造に含める．そして「大腿」を「腿」に,「頚部」を「部」に,「骨折」を「折」によって置き換える．すると「大腿→骨」が「腿→骨」と，単位語間の係り受け関係となるので，これを係り受け構造に含め,「大腿骨」を「骨」と置き換える．これを繰り返していくと，図 4 の係り受け構造を得る．コンピュータ上では図下部のように，1 単位語を「単位語 ID」「単位語」「係り先の単位語 ID」の 3 つ組として表現し，その集合として複合語を表現する.



1	大	2
2	腿	3
3	骨	7
4	頚	5
5	部	3
6	骨	7
7	折	0

図 4　大腿骨頚部骨折の内部構造（係り受け構造）
上部：図表現．矢印は係り受け関係を表し，矢印の向かう先が Head.
下部：図表現と等価な情報を表す，コンピュータ上でのデータ構造．1 つの単位語を「単位語 ID」「単位語」「係り先の単位語 ID」の 3 つ組で表現する．この複合語内に係り先を持たない「折」は仮想的な単位語に係るものとし，その ID を「0」とする．

以上のようにして得られた係り受け構造（図 4）が，本研究で提案する内部構造表現の基本である．内部構造表現の人手による解析手順を以下にまとめる．

文字列表現から内部構造表現への変換方法

1. 以下の処理を，語の分割が出来なくなるまで繰り返す.

文字列 ”$C_1 \ldots C_n$” を ”$C_1 \ldots C_m$” と ”$C_{m+1} \ldots C_n$” に分割する.

分割位置 m を以下の手順で決定する.

1-1. 分割位置の前方探索 m_{prefix}

初期設定：分割候補位置 i を 2 に設定する.

【繰り返し 1】

もし m_{prefix} に値が設定されていたら【繰り返し 1】を終了する.

もし”$C_1 \ldots C_i$”と”$C_{i+1} \ldots C_n$”が共に”$C_1 \ldots C_n$”中での意味を保っている語ならば

m_{prefix} を i に設定する.

そうでなければ

i の値を 1 つ増やす.

もし i の値が n であれば m_{prefix} を n-1 に設定する.

1-2. 分割位置の後方探索 m_{suffix}

初期設定：分割候補位置 i を n-1 に設定する.

【繰り返し 2】

もし m_{suffix} に値が設定されていたら【繰り返し 2】を終了する.

もし”$C_1 \ldots C_i$”と”$C_{i+1} \ldots C_n$”が共に”$C_1 \ldots C_n$”中での意味を保っている語ならば

m_{suffix} を i に設定する.

そうでなければ

i の値を 1 つ減らす.

もし i の値が 0 であれば m_{suffix} を n-1 に設定する.

1-3. 係り受け関係（Head と Modifier）の決定

m_*（*は prefix または suffix を示す）に対して，

1-3-1. is-a 関係

もし”$C_1 \ldots C_n$” is a ”$C_1 \ldots C_{m*}$” が成立するなら

$HEAD_*$ に”$C_1 \ldots C_{m*}$” を設定する．

もし”$C_1 \ldots C_n$” is a ”$C_{m*+1} \ldots C_n$” が成立するなら

$HEAD_*$ に”$C_{m*+1} \ldots C_n$” を設定する．

1-3-2. part-of 関係

もし $HEAD_*$に何も設定されていなかったら

もし”$C_1 \ldots C_n$” is part of ”$C_1 \ldots C_{m*}$” が成立するなら

$HEAD_*$ に”$C_1 \ldots C_{m*}$” を設定する．

もし”$C_1 \ldots C_n$” is part of ”$C_{m*+1} \ldots C_n$” が成立するなら

$HEAD_*$ に”$C_{m*+1} \ldots C_n$” を設定する．

1-3-3. is-a/part-of 関係が無い場合

もし $HEAD_*$に何も設定されていなかったら

$HEAD_*$に”$C_{m*+1} \ldots C_n$” を設定する．

1-4. 分割位置の決定

$HEAD_{prefix}$ と $HEAD_{suffix}$ のうち，文字列長の大きい方の添え字を持つ m_*を分割位置 m とする．

2. 以下の処理を，係り受け関係に現れる語が全て単位語になるまで繰り返す．

係り受け関係にある 2 語が共に単位語である場合について，この関係を係り受け構造に含め, 2 語が合成して出来る語をこの係り受け関係における Head と置き換える．

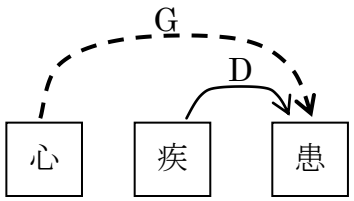
3.1.2. 内部構造の基本表現の拡張

提案する内部構造表現の基本は上記の係り受け表現であるが，2 つの例外現象が存在しており，これらを扱うために表現方法を拡張する．

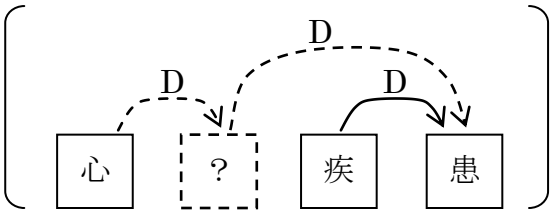
例外現象 1 :「省略現象」

構成要素が結びつく際に，一部の文字（列）が合成語内に現れない場合がある．これを本研究では省略と呼ぶ．逆に，省略の起きた合成語をその構成要素に分割した場合には，合成語の中には現れない文字（列）が構成要素の一部として新しく出現する．例えば「心疾患」は「心の疾患」ではなく「心臓の疾患」であり，「臓」が省略されている．

これを合成語の内部構造に明示的に表現するために，係り受け関係に種類を持たせることとした．これまで使用していた係り受け関係をラベル D（Dependency）で表し，省略を表現する係り受け関係としてラベル G（Generate）を導入した（図 5）．係り受け関係 G は文字を生成する係り受け関係であり，「G で結ばれた二文字の間に何らかの文字が存在し，G の係り元が中間文字に関係 D で係り，中間文字が G の係り先に関係 D で係る」ことを示している．



1	心	3	G
2	疾	3	D
3	患	0	_

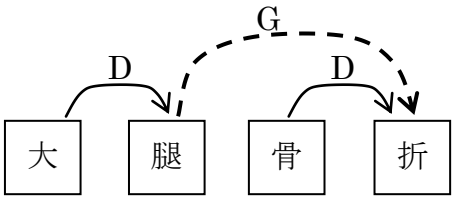


1	心	2	D
2	？	4	D
3	疾	4	D
4	患	0	_

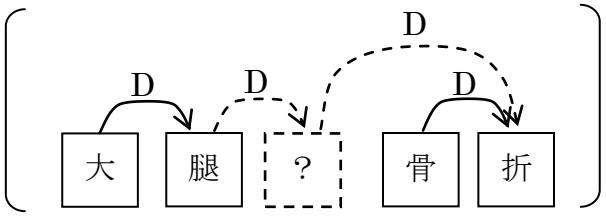
図 5　省略現象の表現方法.
図表現とそれに対応するコンピュータ上での表現を示す．基本表現にラベル情報が付加されている.
左図．係り受け関係の種類として D の他に G を追加した表現.
右図．左図を全て D で表現したもの.「？」は任意の文字を表す.

例外現象２：「縮退現象」

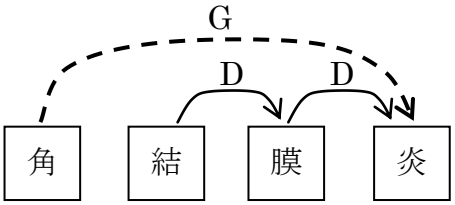
構成要素が結びつく際に，複数構成要素で重複する文字（列）が合成語内では一度だけしか出現しない場合がある．これを本研究では縮退と呼ぶ．縮退の起きた合成語をその構成要素に分割した場合には，合成語内の一部の文字（列）が重複して出現する．例えば，「大腿骨折」では「大腿骨」と「骨折」の共通部分である「骨」が 1 文字消えており，「角結膜炎」では「角膜」と「結膜」の共通部分である「膜」が 1 文字消えている．縮退は省略の一種と考えることが出来，内部構造表現としては省略と同じ枠組みで扱うことが可能である（図 6）.



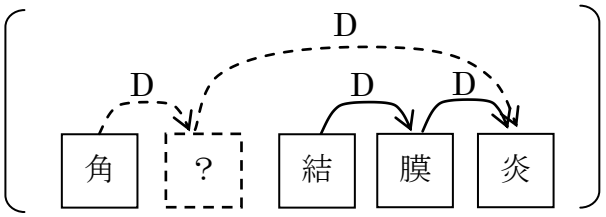
1	大	2	D
2	腿	4	G
3	骨	4	D
4	折	0	__



1	大	2	D
2	腿	3	G
3	骨	5	D
4	骨	5	D
5	折	0	__



1	角	4	G
2	結	3	D
3	膜	4	D
4	炎	0	__



1	角	2	D
2	?	5	D
3	結	4	D
4	膜	5	D
5	炎	0	__

図 6　縮退現象の表現方法.
左図. 係り受け関係の種類として D の他に G を追加した表現.
右図. 左図を全て D で表現したもの.「?」は任意の文字を表す.

省略・縮退現象を表現するために係り受けラベルを導入した場合，内部構造表現の付与方法は，1）語の分割の際に省略・縮退を復元させること，2）その結果得られる前述のラベルなし内部構造表現からラベルを導入した内部構造表現への変換，の 2 点の変更が加わり，以下のようになる．拡張前の方法との主な変更点を下線で示す．

文字列表現から内部構造表現への変換方法の拡張

１．以下の処理を，語の分割が出来なくなるまで繰り返す．

文字列 ”C_1…C_n” を STRING1 と STRING2 に分割する．

ただし，STRING1 は必ず文字”C_1”を，STRING2 は必ず文字”C_n”を含む．また STRING1 と STRING2 を連結させた文字列”C'_1…C'_{n+m}” (m は非負整数)は，文字の n 組(C_1…C_n)を，順序を保って内包する．

分割 M を以下の手順で決定する．

分割の前方探索 M_{prefix}

初期設定：STRING1 の長さが最も短くなるような分割 I を設定する．

【繰り返し 1】

もし M_{prefix} に値が設定されていたら【繰り返し 1】を終了する．

もし STRING1 と STRING2 が共に”C_1…C_n”中での意味を保っている語ならば

M_{prefix} を I に設定する．

そうでなければ

I の値を，STRING1 の長さを 1 つ増やしたものに更新する．

もし STRING1 の長さが n であれば M_{prefix} を 1 つ前の I に設定する．

分割位置の後方探索 m_{suffix}

初期設定：STRING2 の長さが最も短くなるような分割 I を設定する.

【繰り返し 2】

もし M_{suffix} に値が設定されていたら【繰り返し 1】を終了する.

もし STRING1 と STRING2 が共に"$C_1 \cdots C_n$"中での意味を保っている語ならば

M_{suffix} を I に設定する.

そうでなければ

I の値を，STRING1 の長さを 1 つ減らしたものに更新する.

もし STRING1 の長さが 0 であれば M_{suffix} を 1 つ前の I に設定する.

Head の決定

M_*（*は prefix または suffix を示す）に対して,

is-a 関係

もし「"$C_1 \ldots C_n$" is a STRING1」が成立するなら

$HEAD_*$ に STRING1 を設定する

もし「"$C_1 \ldots C_n$" is a STRING2」が成立するなら

$HEAD_*$ に STRING2 を設定する

part-of 関係

もし $HEAD_*$に何も設定されていなかったら

もし「"$C_1 \ldots C_n$" is part of STRING1」が成立するなら

$HEAD_*$ に STRING1 を設定する

もし「"$C_1 \ldots C_n$" is part of STRING2 」が成立するなら

$HEAD_*$ に STRING2 を設定する

is-a/part-of 関係が無い場合

もし HEAD*に何も設定されていなかったら

HEAD*に STRING2 を設定する.

分割位置の決定

$HEAD_{prefix}$ と $HEAD_{suffix}$ のうち, 長い方の HEAD*に対応する M*を分割 M とする.

ラベルの付与

もし STRING1 と STRING2 を連結した文字列"$C'_1 \ldots C'_{n+m}$"が"$C_1 \ldots C_n$"に等しければ

係り受けラベルを D に設定する.

そうでなければ係り受けラベルを G に設定し, STRING1 の末尾 m 文字を削除する.

２．以下の処理を，係り受け関係に現れる語が全て単位語になるまで繰り返す.

係り受け関係にある 2 語が共に単位語である場合について, この関係を係り受け構造に含め，2 語が合成して出来る語をこの係り受け関係における Head と置き換える.

3.2. 自動 ICD コーディング手法への応用

上述した内部構造表現法の医療における有用性を検討するため，既存の自動 ICD コーディング手法において内部構造表現を利用した時のコーディング精度の変化を検討した．

3.2.1. ICD コーディング

「疾病及び関連保健問題の国際統計分類（International Classification of Diseases and Related Health Problems: ICD)」[23]は世界保健機関が公表する分類であり，死因や疾病の統計に利用される．最新の分類は 1990 年に採択された ICD-10 と呼ばれる．現在我が国では 2003 年度版の ICD-10 に準拠して作成された「疾病，障害及び死因分類」が厚生労働省から公表されており，各種統計調査や診療録管理等はこの分類に基づいて行われている．本稿では「疾病，障害及び死因分類」を「ICD 分類」，これに収載された分類コードを「ICD コード」，疾患名に対して ICD コードを付与することを「ICD コーディング」と呼ぶ．

医療機関では各診断に対して ICD コーディングが行われる．しかしコーディング作業は煩雑であり，また作業者の能力も多様であるため，コーディングの質は保証されない．現在，作業を補助するものとして標準病名マスター[24]が公開されている．標準病名マスターの根幹は ICD コードを予め付与した基本病名のリストであるが，これだけでは十分とは言えない．なぜならば入力疾患名は一般に自由入力であるために表記が統一されておらず，そのすべてを網羅したリストを予め用意する（Pre-coordination と呼ばれる）ことができないからである．

ここで，同一概念に対する異表記（表記揺れ）について述べる．ICD コーディングは，コーディング対象の疾患名が表す概念と同一あるいは上位概念を表すICD コードを選択する作業であり，文字列から概念へのマッピングである．ところでマッピング先は概念であるが，これを表現するのは言語であるため，結局，ICD コーディングとは「入力文字列と同じ概念を表す文字列を, 有限個の文字列集合から選択するタスク」と言うことができる．これは有限個の文字列集合から入力文字列と表記揺れ関係にある文字列を選択することに等しい．

表記揺れには以下の三つのパターンが考えられる．

1. 翻字

例：アスペルガー / アスペルガ

発音とその表記によるパターンであり，その多くが外来語である．バリエーションが非常に多く，あらかじめその全てを記述しておくのは困難であるが，外部知識を必要とせず文字列のみから判断が可能である．翻字による表記揺れの解消は自然言語処理学分野で様々な手法が提案されている[25,26]．

2. 同義語

例：コーツ病 / 滲出性網膜炎

オントロジ等の外部知識が必要となるパターンである．

3. 修飾語の順序や有無

例：急性 A 型肝炎 / A 型急性肝炎

大腿骨頚部骨折 / 大腿骨頚部開放骨折 （同一 ICD コードを持つ）

非常にバリエーションの多いパターンであり，これに関する外部知識の作成は非常にコストがかかるが，パターン 2 と異なり外部知識が無くともある程度の推測が可能である．

標準病名マスターではパターン 1 については対応せず（『標準』病名としてパターン 1 のような表記揺れは認めるべきではなく，また人手によるコーディング作業では問題となりにくいパターンである），パターン 2 については同一 ICD コードに対して同義語を収載することで対応している．本研究で問題とするのはパターン 3 である．

標準病名マスターは基本病名のバリエーションの少なさを補う手段として修飾語テーブルを用意している．ICD コードが既知である基本病名に対して任意個の修飾語の付加を許容することで表現力を拡充しているのである（Post-coordination）．ところが，このようにして作られた合成疾患名の ICD コードとして元の疾患名に付与されていた ICD コードがそのまま使われるが，必ずしもそれは正しくない．多様な入力に対応するための手段がコーディング結果の質に対して副作用を及ぼし，ICD コーディングの質を下げる結果となっている．この Post-coordination によって作られた疾患名における表記揺れは前述の表記揺れパターン 3 に相当するものである．

ところで，2 つの語が与えられた時，両者がパターン 3 の表記揺れ関係にあり同一概念

を表していると判断するためには，表層文字列だけではなく構成要素間の関係を考慮にいれる必要がある．例えば「急性 A 型肝炎」と「A 型急性肝炎」を同じであると判断するためには，両者において「急性」と「A 型」は共に「肝炎」を修飾することを知らなければならない．もし「急性」が「A 型」を修飾すると捉えると，「急性 A 型肝炎」は「『急性 A 型ウイルス』というウイルスによる肝炎である」という誤った解釈が成り立つからである．

修飾・被修飾の関係を正しく捉えるということは内部構造を知ることに等しい．従ってパターン 3 の表記揺れを解消するうえでは本研究で提案した内部構造表現の利用が有効であると考えられる．

自動 ICD コーディングの試みはこれまでにも多く行われており，ICD コードが既知である用例との類似度を用いる手法[27-29]と，ICD コーディングの知識を記述しそれを用いる手法[30-32]に大別される．前者は実装が簡単であるが用例が大量に必要であるという難点を持っており，後者は記述にかかるコストが高いという難点がある．その中でも今井ら[30]は後者の手法を ICD 分類全体に適用したが，入力疾患名の解析を形態素解析（構成要素への分割）のレベルでしか行っていない．

そこで，本研究が提案する内部構造表現法を利用することで，今井らの手法のコーディングの精度に与える影響について検討を行うこととした．なお，本研究における「自動 ICD コーディング」は，人による自由入力された疾患名に正確な ICD コードを付与することを目的としたものであり，診療録等の情報は用いず，また入力疾患名を統制するものではない．

3.2.2. コーディング手法

ベースラインとなる手法

まず本研究の提案手法の比較対象である今井らの先行研究[30]の手法について述べる．今井らの手法は，ICD 分類を人手で構造化し[40]，この情報と疾患名の部分文字列の照合を行ったものである．以下，この手法を「既存手法」と呼び，以下に概略を説明する．

1. 構造化 ICD の作成

今井らは，ICD-10 分類のうち，精神疾患・症状/所見・外因などを除く主要な 15 の章(1,2,3,4,6,7,8,9,10,11,12,13,14,17,19)を人手で構造化した. 構造化は以下の 2 つの段階を経て行われた.

Step1) ICD-10 分類の各エントリ（大分類，中分類，細分類）に相当する概念が持つ意味関係を人手で記述した. 作業は診療情報管理士並びにそれに準ずる能力を持った 20 名によって行われ，結果を別の 2 名によって複数回精査し，随時修正を行った．各エントリは，ICD コード，二重分類情報，内容説明・限定・指示情報，「～不明・～以外」関連の詳細情報，意味関係情報の 6 項目が付与されている．意味関係情報は「関係（以下< >)」と「関係対象概念（以下[ ])」の組で表現され，例えば顆粒性結膜炎に対しては <主病態>[結膜炎]， <症状･所見>[顆粒状変性] 等が付与されている．また各関係対象概念は，同時に自然言語上の表記ラベル（以下 ""）を 1 つ以上保持しており, 例えば概念 [甲状腺ホルモン調節機能正常] は "甲状腺ホルモン調節機能正常" だけでなく，"非中毒性" という表記ラベルも持っている.

Step2）上記ステップではICDの各分類コードの見出し語をベースに関係対象概念の切り出しを行ったが，これだけでは各コードに分類される疾患が持つ意味関係を十分カバーしきれないため，標準病名マスターに収載された疾患名から切り出された概念を用いて,以下のような意味関係と表記ラベルの拡充を行った.

1）関係対象概念の表記ラベルの追加

例）［ 好酸球 ］に対する　"エオジン細胞"

2）関係対象概念の下位概念とその表記ラベルの追加

例）［ 乳房 ］に対する　［ 乳頭 ］（"乳頭", "乳嘴部"）

3）意味関係情報の追加

例）M254【関節浸出液貯留】 =>　<症状・所見>　［ 腫脹 ］（"腫脹"）

2.　コーディング方法

まず入力疾患名を自作の解析器を用いて分割して部分語の集合を得る．この部分語の集合と構造化 ICD の意味関係情報を比較し，ある ICD カテゴリの 1）関係対象概念の表記ラベル，2）関係対象概念の下位概念の表記ラベル，3）上位カテゴリの関係対象概念の表記ラベル，だけで構成されているような ICD コードを入力疾患名に付与する．この基準を満たす ICD コードが存在しなかった場合は，半数以上の部分語についてこの基準を満たし，かつ最も被覆度の高い ICD コードが候補として出力される．

3.　評価実験

全国病院で実際に入力された疾患名から標準病名マスターに収載されていない

1211 疾患名をランダムに選び，人手により正解 ICD コードを付与した．このデータに対し，上記の自動 ICD コーディング手法を適用したところ，正しい ICD コードを付与されたのは 747 疾患名（61.7%）であったと報告している．

以上の既存手法では，構造化 ICD には意味関係情報として係り受け情報に相当する情報が存在するが，解析対象となる入力疾患名の処理は分割のみにとどまっており，係り受け情報は考慮していない．係り受け情報を考慮に入れないと，例えば「嚢胞 | 腎」と「腎 | 嚢胞」は共に「腎」と「嚢胞」の組として表現され区別ができず，「大腿骨 | 頚部 | 骨折」を「大腿骨と頚部の骨折」と捉えることもできてしまう．そこで係り受け情報に準ずる情報として，入力疾患名の内部構造表現を利用することにより，より精度向上が期待できると考えた．ここで，構造化 ICD は内部構造表現に相当する文字単位の係り受け情報を持たず，上記の手法をそのまま適用することはできないため，本研究では既存手法を基盤として次項で述べるような複合語の内部構造の係り受け情報を付加した自動ICD コーディング手法を提案し，既存手法との精度比較を行った．

本研究の提案する自動 ICD コーディング手法

「内部構造表現→ICD コード」の形のコーディングルールを用意し，入力疾患名の内部構造表現がルール左辺を含む場合にルールを適用することとした.

入力疾患名の内部構造表現に対して複数のルールが適用可能である場合は，ルール左辺の内部構造に現れる文字数が最も多いものを選択する．これは，(1)文字が増えるほど語が示す概念の粒度が細かくなる，(2)ICD 分類における概念は粒度が不均一である，(3)該当 ICD コードが複数あった場合は最も細かい粒度の ICD コードに分類するべきである，という仮定に基づいた方針である. 例えばコーディングルールとして「大腿骨頚部骨折→S72.0」「大腿骨骨折→S72.9」が与えられた場合,「大腿骨頚部骨折」は両ルールに適用できるが，上述の方針に従い，文字数が多い前者のみを適用させることとする.

提案手法と既存手法の相違点は，1）構成要素である語（文字列）が持つ情報を使用せず文字の繋がり（係り受け）の情報のみを使用すること，2）入力疾患名の内部構造を考慮に入れること，3）構造化 ICD に意味関係情報として含まれる異表記を利用しないこと，の 3 点である．既存手法に対して提案手法が有利な点は 2，不利な点は 3 である.

ルールの作成

理想的には，ルール左辺に現れ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が，本研究では ICD-10 分類の細分類見出し文字列を内部構造表現で記述することでルール左辺とした．

ICD-10 の分類見出しは文字列表現としてみると名詞列である複合名詞ではないものや，指し示す概念が曖昧なものがある．ここでは，このような場合にどのようにルールを記述すればよいかについて述べる．

表層文字列によって分類すると，表 1 のようになる．ここで，機能表現とは「による」「の」「を含む」のように，概念の内容ではなく文や語を構成する機能を持つ表現を指す．

表 1　ICD-10 の見出しの分類

		分類情報	
		明示されている	明示されていない
機能表現	含む	レンサ球菌による咽頭炎	その他の急性副鼻腔炎 急性副鼻腔炎，詳細不明 その他の明示された病原体による急性咽頭炎
	含まない	急性上顎洞炎	（該当なし）

まず，機能表現を含む場合は，機能表現部分をどのように扱うかが問題となる．方法としては以下の三つが考えられる．

(1) 何もしない

(2) 機能語を取り除く（レンサ球菌による咽頭炎 → レンサ球菌咽頭炎）

(3) 機能語を置き換える（レンサ球菌による咽頭炎 → レンサ球菌【性】咽頭炎）

内部構造の単位語は意味を持つものの方が望ましいことを考慮すると，(3)が最善策となる．しかしそのためには機能表現範囲を特定し，何らかの語と置き換えるという処理が別途必要になり，また置き換え先の語として何を選択するべきかという問題が新たに発生する．本研究では，機能表現はその前後にある語を結びつける役割を果たしており，語全体の内容に対する比重は軽いことから，(2)を採用した．

また，見出し中の＜＞による異表記，「および」等による並列は全て展開して記述した．例えば「肺胞性及び肺胞周囲性病態」は「肺胞性病態」「肺胞周囲性病態」，「成人呼吸窮<促>迫症候群<ＡＲＤＳ>」は「成人呼吸窮迫症候群」「成人呼吸促迫症候群」「ＡＲＤＳ」と分割し，それぞれの内部構造表現のいずれかを包含していれば当該コードを付与するというルールを作成した．

次に，分類情報が明示されていない場合であるが，これは更に二つに分割することができる．1) 当該見出しには分類情報が明示されていないが，他の見出しと併せることで暗に分類基準が示されている場合と，2) 他の見出しと併せても分類基準が不明な場合である．

前者について，例えば「その他の明示された病原体による急性咽頭炎[J02.8]」はそれだけでは「その他の明示された病原体」が何であるのかわからないが，同じ J02 の他分類「レ

ンサ球菌による咽頭炎[J02.0]」を考慮すれば，「レンサ球菌以外の病原体」を示していることがわかる．

また「肺炎を伴わない肺膿瘍[J85.2]」の場合，「肺炎を伴わない」という情報は普通疾患名の中では明示されないが，「肺炎を伴う肺膿瘍[J85.1]」に分類される疾患名では「肺炎を伴う」ことが明示されるので，明示されていなければ「肺炎を伴わない」と判断するという基準が成り立つ．従ってこの場合は「肺膿瘍」のみをルールに記述すればよい（図7）．これは概念を明確に記述するという意味では不十分であるが，デフォルト値として否定表現を採用することは暗黙知を仮定していることになり，このような「表現しない」ことに意味を持たせるのは一つの方法であると考えられる．

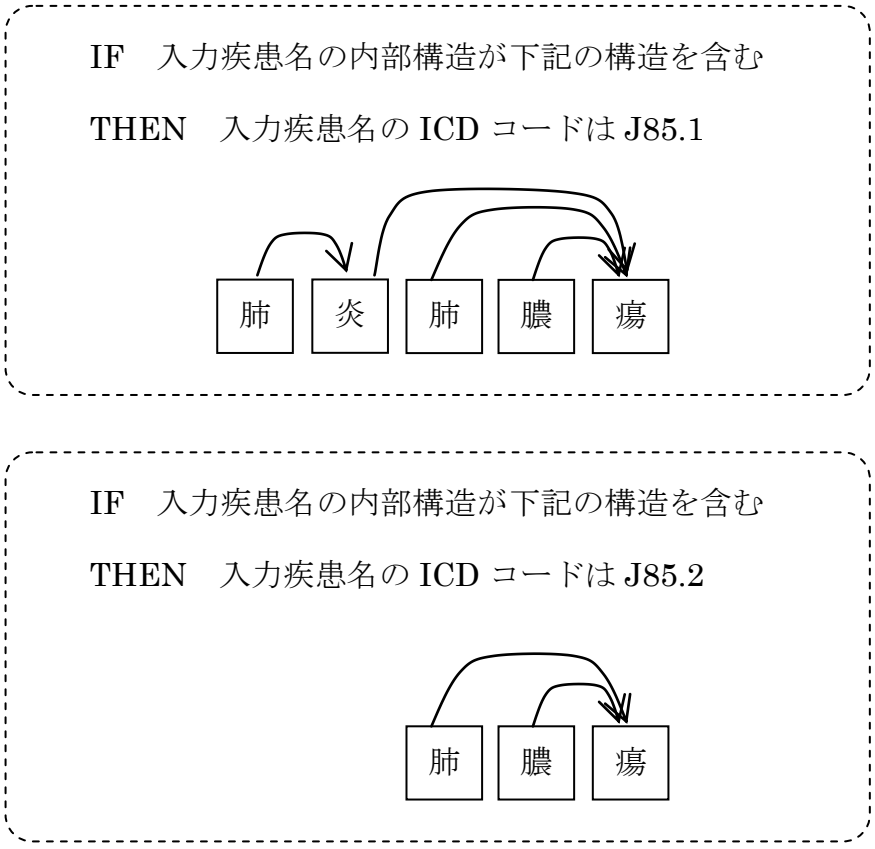
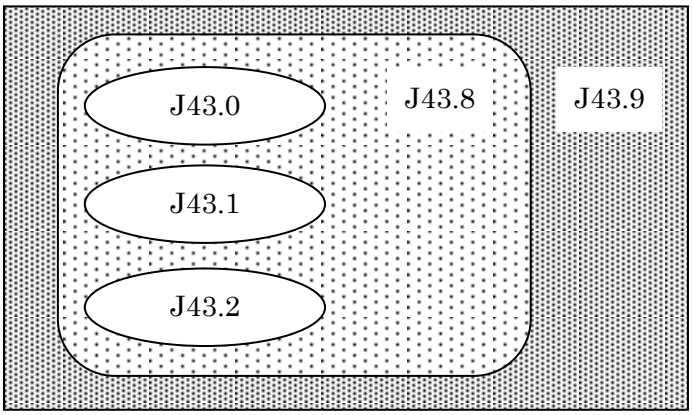


図 7　【肺炎を伴う肺膿瘍】と【肺炎を伴わない肺膿瘍】のコーディングルール

後者の典型例は「その他の<疾患>」と「<疾患>，詳細不明」のペアであり，両者の違いは見出し文字列からの判断は不可能であり，用例を参照して類推しなければならない（図 8）．これは ICD が死因統計を目的とした分類であり，疾患概念の分類基準としては恣意的なものであることに起因する．

ただし，「その他の」が修飾する先によっては分別可能な場合がある．例えば「その他のウイルス肺炎[J12.8]」」の場合，「その他の」は「ウイルス」を修飾しており，ウイルスは一般的に「～ウイルス」の形で表現されるため，図 9 のようなルールが有効である．



J43.0　マクロード症候群
J43.1　汎小葉性肺気腫
J43.2　中心小葉性肺気腫
J43.8　その他の肺気腫
J43.9　肺気腫，詳細不明

図 8　ICD 分類の恣意性

J43.8 は J43.0～J43.2 ではないという集合，J43.9 は J43.0～J43.2 および J43.8 ではないという集合．両者の差は「その他」と「詳細」の意味に依存するが，これは分類体系の作り方に依存して決まるものであり，疾患概念の本質ではない．

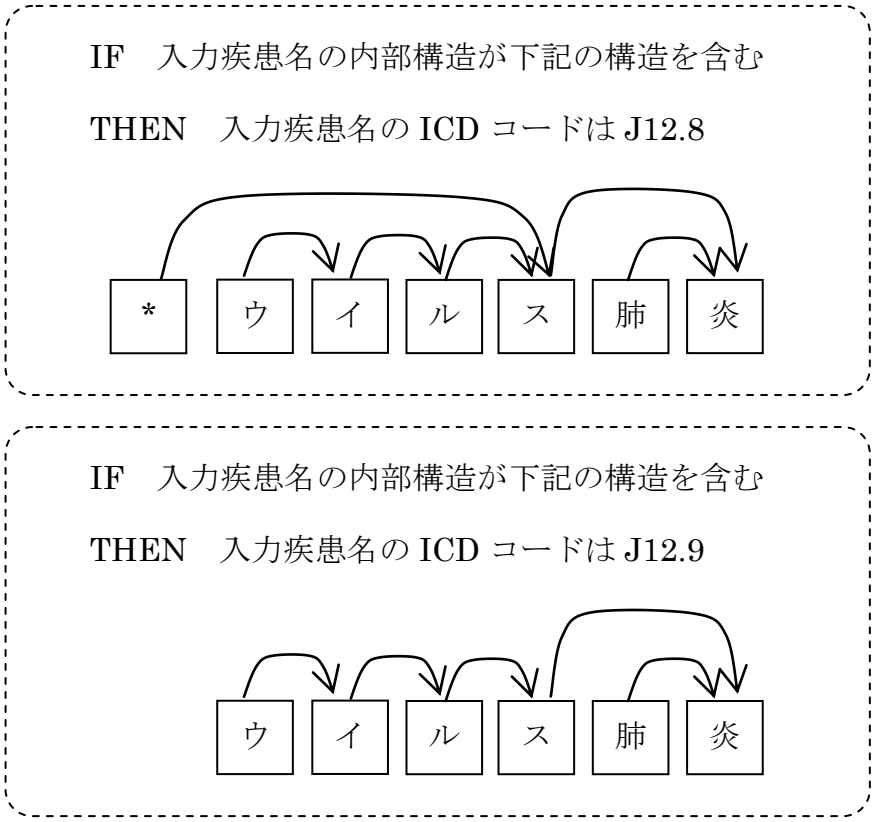


図 9　【その他のウイルス肺炎】と【ウイルス肺炎，詳細不明】のコーディングルール

提案手法による表記揺れ解消過程

以上の提案手法は表記揺れパターン 3「修飾語の順序・有無」の解消を期待するものであるが，その解消過程を以下に示す．

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（図 10）．

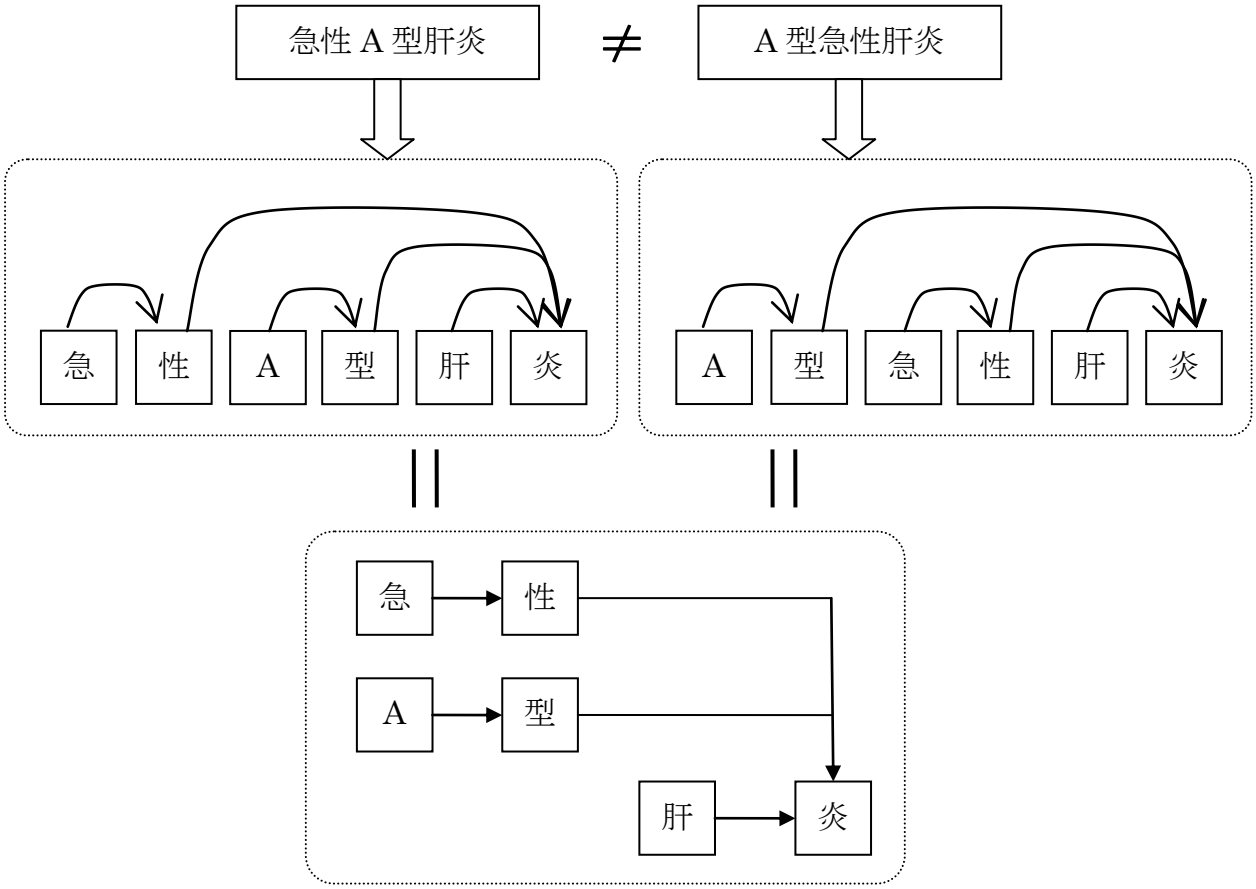


図 10　修飾語の順序による表記揺れの解消
同じ概念を表す「急性 A 型肝炎」と「A 型急性肝炎」は，文字列表現は異なるが，内部構造表現は等しい．内部構造表現に変換することで，修飾語の順序の違いによる表記揺れを解消できる．

次に，内部構造表現で表した入力疾患名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（図 11）.

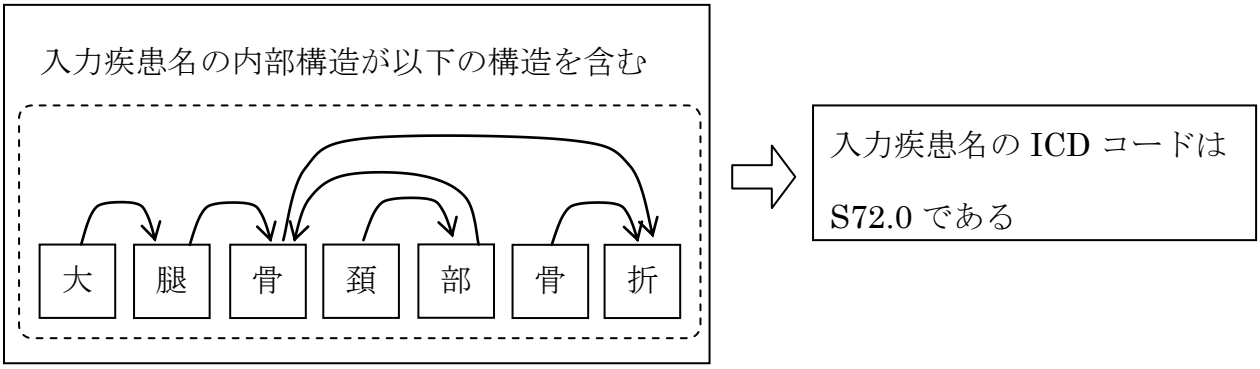


図 11　S72.0【大腿骨頚部骨折】のコーディングルール

入力疾患名の内部構造表現に対してコーディングルールを適用すると，入力疾患名内にある修飾語のうち，ICD コーディングには不必要なものが無視される（図 12）. すなわち，表記揺れパターン 3 のうち「修飾語の有無」によるものが解消される.

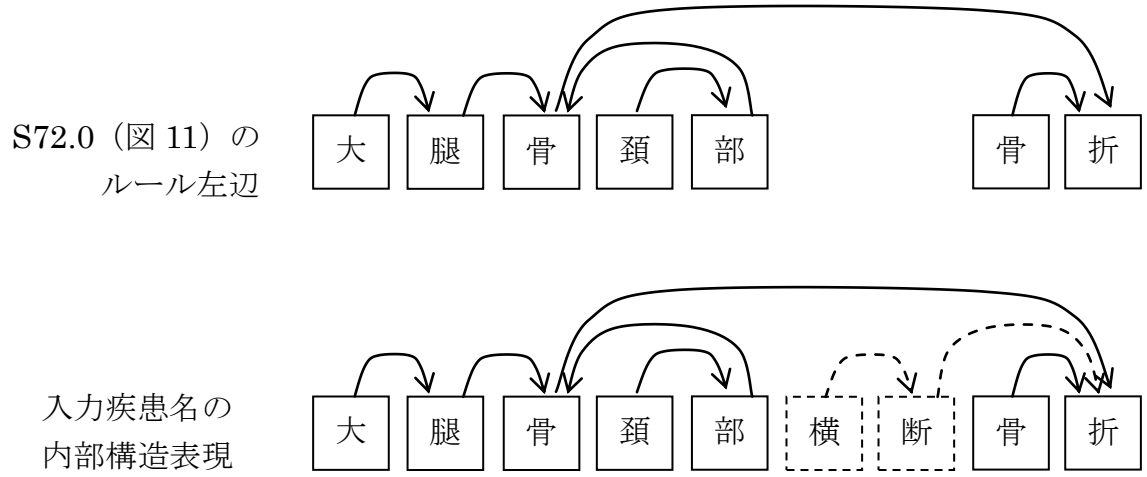


図 12　修飾語の有無による表記揺れの解消
実線で 2 つの内部構造表現の共通部分を示す．入力疾患がルールに記述された内部構造を包含するという条件により，「大腿骨頚部横断骨折」には図 7 のルールが適用される．この時，S72.0 では不必要な修飾語「横断」が無視され，表記揺れパターン 3 のうち「修飾語の有無」に関する部分が解消される．

もし内部構造を考慮せず文字列のみの照合を行った場合，入力疾患名「大腿骨頚部横断骨折」では S72.0 の分類見出し「大腿骨頚部骨折」の中に「横断」が挿入されて「大腿骨頚部」と「骨折」が分断されているため，両者は一致せずコーディングは失敗する．

3.2.3. 提案手法の評価方法

前提条件

この提案手法の性能に影響を与えるものは以下の 2 点である.

1) 文字列表現で与えられた入力疾患名を内部構造表現に正しく変換する性能

2) コーディング手法の性能

ここで，本研究で主眼を置くのは疾患名の内部構造表現法の提案であり，自動 ICD コーディングは内部構造表現の有用性評価のために取り上げた応用例であることを明記しておく. 従って自動 ICD コーディングの実験においては，前提として，入力疾患名は既に正しい内部構造表現として与えられることとした.

材料

全国病院で実際に入力された疾患名から標準病名マスターに収載されていない 1211 疾患名をランダムに選び，人手により正解 ICD コードが付与されたデータを用いた．これは既存手法で実験データとして使われたものと同じものである.

この中で，既存手法でコーディングできなかった 508 疾患名から 100 疾患名をランダムに選択し，提案手法でコーディング可能か否かの実験を行った．実験は医療情報学に数年関わった非医療従事者 1 名（著者）が行った.

既存手法でコーディング可能であった疾患名を実験の対象としていない理由について述べる．本研究における自動 ICD コーディング手法は，既存手法と対立するものではなく，

そのコーディング性能を改善するためのものである．入力疾患名の内部構造表現として与えられたとしても，元の文字列表現を用いれば既存手法によるコーディング結果の正誤は変化しない．あるいは既存手法における入力疾患名の処理を，文字列すなわち「先頭文字から末尾文字まで順に連なる形の木」の分割ではなく，内部構造表現すなわち「より複雑な文字の木」の分割をするとすれば，やはり結果は変わらないと考えられる．従って，内部構造表現の有用性評価としては既存手法で正しくコーディングできなかった疾患名のみを対象とすることは妥当な方法である．

評価方法

評価方法として，既存手法と同様に精度として正解率，すなわち「正しいICDコードが付与された疾患名の数 / 実験の対象となった疾患名の数」を計算する．前述したとおり，実験対象となる疾患名は既存手法による自動ICDコーディングで正しくコーディングできなかった疾患名であり，既存手法で正しくコーディングできた疾患名集合は内部構造表現を導入しても変化しないため，本研究で計算される精度は既存手法の精度 61.7%からの増加分を意味する．

3.3. 内部構造解析

前節で述べた自動 ICD コーディング手法は，コーディング対象となる入力疾患名の内部構造表現が与えられているという仮定をおいた．しかし実際には疾患名は文字列表現として存在するものであるため，文字列として入力される疾患名を自動解析し内部構造を与える解析器が必要である．本節では汎用的な係り受け解析器を用いた内部構造解析について述べる．

3.3.1. 内部構造解析の方法

本研究で提案した内部構造表現は一般に使用される文の係り受け関係と同じ枠組みによる表現であるため，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[5,7,33-37]．アルゴリズムは，係り受け関係の種類（ラベル）無/有，決定的/非決定的のように分類される．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解析器 MaltParser[38,39]を用いた．

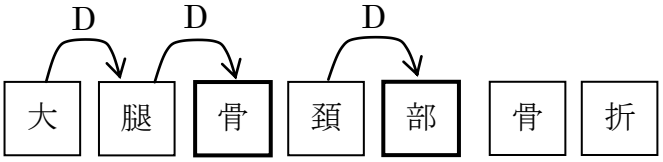
MaltParser は機械学習機能を備えた汎用的な係り受け解析器であり，要件を満たした係り受け構造であればユーザ定義のものでも扱うことができる．解析において，MaltParser は解析対象となる単位語列を前から順に見ていき，次の単位語との間に係り受け関係が存在するかを判断する．この判断は分類タスクであり，分類タスクに対してはさまざまな機械学習手法が提案されている．MaltParser は機械学習手法として非常に性能が良いことが

理論的にも実験的にも示されているサポートベクターマシン（SVM）を用いている．学習を行うためには係り受け関係の成立する/しない単位語対を学習用データとして用意する必要があるが, MaltParser には正しい係り受け構造を付与したデータを用いてこの学習を行うという機能が実装されている．一度学習を行うと，MaltParser 内部の SVM は単位語対の間に係り受け関係が存在するかどうかの判断機能を獲得し，この SVM を用いることで MaltParser は係り受け構造が未知であるデータを解析できるようになる．学習および解析の際には，単位語そのもののみでなく，単位語の特徴を表す様々な情報（素性）が利用可能ある．本研究で使用した素性を表 2 に，素性の具体例を図 13 に示す.

表 2　内部構造解析で使用した素性

Stack，Input は MaltParser の内部で使用されるデータ構造．Stack には解析済み部分，Input には未解析部分を格納しており，MaltParser は Stack[0]と Input[0]の間に係り受け関係が存在するか否かを判定する．MaltParser は係り受け関係の有無の判定のために二値分類器（サポートベクターマシン（SVM））を使用している．

分類	ID	説明	データ型
文字素性	1	文字	STRING
	2	文字種	STRING
	3	文字の位置	STRING
辞書素性	4	この語の部分文字列でこの文字を最後とする辞書掲載語が存在する	BOOLEAN
	5	4 の辞書掲載語の長さ	INTEGER
	6	4 の辞書掲載語の MeSH カテゴリ（1 桁+3 桁+全て）	STRING
	7	この語の部分文字列でこの文字を最後とする接尾辞が存在する	BOOLEAN
	8	この語の部分文字列でこの文字を含む辞書掲載語が存在する	BOOLEAN
	9	この文字と n 文字先をつなげた 2 文字語が辞書掲載語が存在する	BOOLEAN
	10	9 での n	INTEGER
係り受けラベル	11	Stack[0]，Stack[1]，Input[0]，Input[1]の係り受けラベル	
	12	Stack[0]の最左/右 modifier の係り受けラベル	
	13	Input[0]の最左 modifier の係り受けラベル	
	14	Stack[0]の左隣の文字の係り受けラベル	



「骨」に関する素性
1: “骨”
2: 漢字
3: 3
4: TRUE (※”骨”)
5: 1
6: A02.835.232
7: TRUE(※部)
8: TRUE(※大腿骨頚部)
9: TRUE(※骨折)
10: 4

「部」に関する素性
1: “部”
2: 漢字
3: 5
4: FALSE
5: (null)
6: (null)
7: TRUE(※部)
8: TRUE(※大腿骨頚部骨折)
9: FALSE
10: (null)

図 13　MaltParser に与えられる文字及び辞書素性.
数字は表 2 の ID と対応する.
図中の矢印とラベルが MaltParser の解析によって既に付与されている時,「骨」と「部」の間の係り受け関係を同定するために，解析器に与えられる情報．この情報をベクトルとして与えることで，MaltParser 内部の SVM が係り受け関係の有無を判定する.

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.

3.3.2. 内部構造解析の精度評価方法

MaltParser による内部構造解析の性能評価のため，以下の設定で評価実験を行った．

- 比較手法

  「全ての文字について, 次の文字を係り先とし, 係り受け関係のラベルは D とする」を提案手法に対する比較手法として採用した．これは係り受け解析の性能評価で一般的に使われる比較手法である．

- 評価指標

  一般的な係り受け解析の評価方法にのっとり，以下の二つを指標とした．

  1）C-ACC：文字対係り受けの解析性能

  2）W-ACC：語（文字対係り受けの集合）の解析性能

  解析性能の指標としては,「正しく解析した対象の数 / 解析した対象の数」で定義される精度を用いた．例えば図 14 の場合，C-ACC=2/3，W-ACC=0/1 となる．

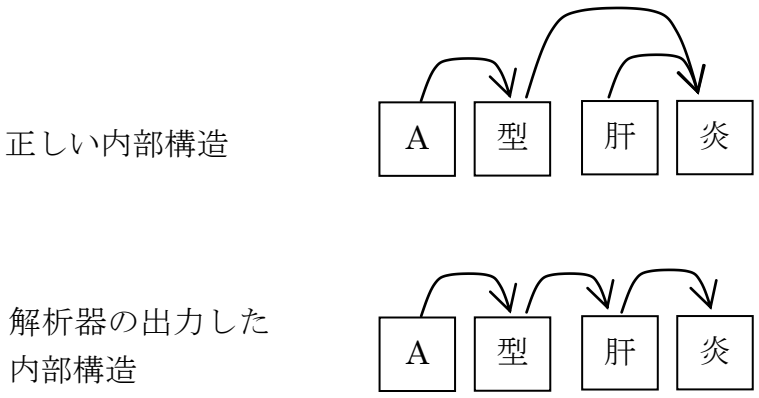


図 14．内部構造解析の評価指標の計算方法
2 つの内部構造を比較すると,「型→炎」と「型→肝」の矢印が違い，他 2 つの矢印は同じであるので, 文字対に対する精度 C_ACC は 2/3 である．一方, 語の精度 W_ACC では 1 語の中の係り受け関係全てが正解でなければならないので，この場合 W_ACC=0 (0/1)となる．

なお，臨床検査や情報検索の性能評価指標として利用される感度・特異度や適合率・再現率といった指標は使用しない．なぜならば，臨床検査や情報検索が解析対象を陽性・陰性，あるいは関連の有無という 2 クラスのいずれかに分類するタスクであるのに対し，係り受け解析は解析対象（単位語）に対して修飾先となる単位語を 1 つ決めるタスクであるためである．また，母数をあるクラスに限定する感度・特異度に対し，上記の精度は全解析対象を母数としており，情報は集約されていることによる．

- 材料

標準病名マスターの収載語のうち，ICD コードが C，E，G，H，K，L の軸からそれぞれ 114 語，96 語，101 語，125 語，123 語，137 語について，人手で内部構造を記述した．このデータを用いて内部構造解析器の学習および精度評価を 5 分割交差検定により行った．

4. 結果

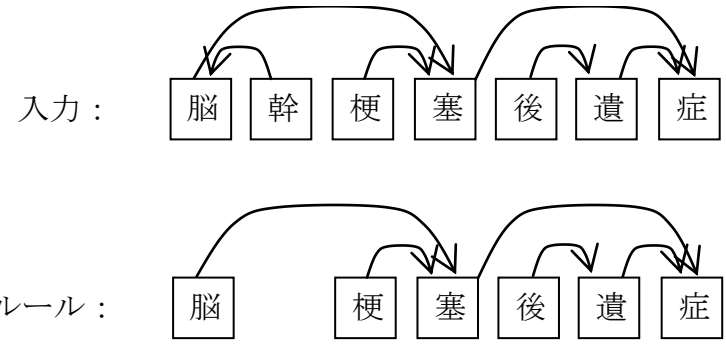
4.1. 自動 ICD コーディング

既存手法でコーディング不可能であった 100 例について調査した結果，提案手法でコーディング可能な語は 8 語あった．正しくコーディングできた要因は(1)内部構造表現，(2)ルールと入力の照合方法のいずれかである．本研究で焦点を当てるのは自動 ICD コーディングの精度よりも内部構造表現を利用する意義であり，これは(1)に当たる．該当する事例は以下の 3 例である．

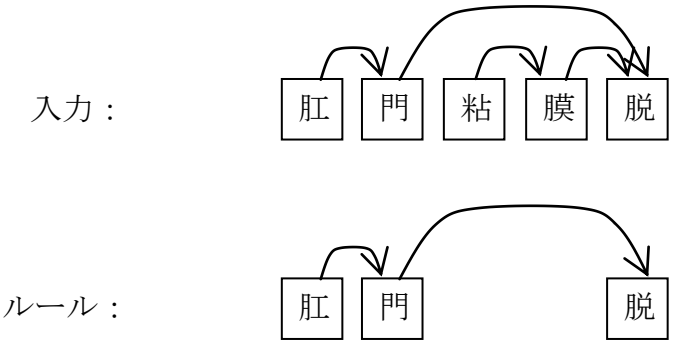
- 脳幹梗塞後遺症（I69.3 【脳梗塞の続発・後遺症】）
- 肛門粘膜脱（K62.2 【肛門脱<脱肛>】）
- 膵頭部のう胞（K86.2【膵のう<嚢>胞】）

これらの入力およびルール左辺における内部構造表現を図 15 に示す．

脳幹梗塞後遺症（I69.3 【脳梗塞の続発・後遺症】）



肛門粘膜脱（K62.2 【肛門脱<脱肛>】）



膵頭部のう胞（K86.2【膵のう<嚢>胞】）

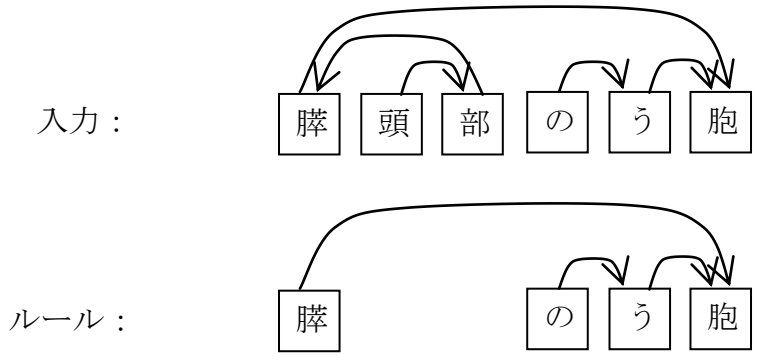


図 15　要因 1（内部構造情報）によりコーディング可能となった事例

(2)ルールと入力の照合方法がコーディング成功の要因となったのは以下の 4 例であった.

- 手汗疱（L30.1 【異汗症[汗疱]】）
- 手骨折（S62.8【手首及び手のその他及び部位不明の骨折】）
- 腕筋肉痛（M79.1【筋(肉)痛】）
- イレウス再発（K56.7【イレウス，詳細不明】）

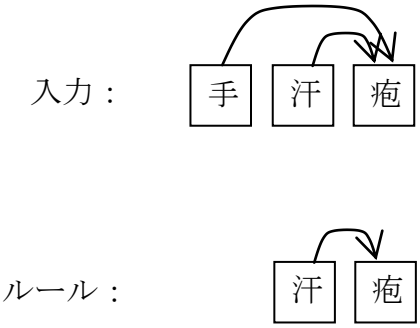
これらの入力およびルール左辺における内部構造表現を図 16 に示す. いずれの入力疾患名も，ICD 分類見出しとの間に情報量の差異がある.「手汗疱」では見出しに「手」の情報が含まれておらず,「手骨折」では見出しに含まれる「その他及び部位不明」の情報が入力疾患名に含まれていない. 本研究で定めた照合方法では「書かれていないことは無視する」という方針を取っているが，既存手法は「全ての構成要素について同値もしくは上位下位関係がある」という本研究よりも厳しい条件を使用しているために，このようなコーディング可/不可の差が出てきたのである．従ってこれら 4 例については提案した内部構造表現が寄与したものではない.

上記二つの要因が複合した例として以下の一例があった.

- 第 5 中手骨頚部骨折（S62.3【その他の中手骨骨折】）

入力およびルール左辺における内部構造表現を図 17 に示す．この語の中で,「頚部」は「中手骨」を修飾しており要因(1)に当たる.「第 5」はルールには表れておらず，要因(2)に当たる部分である（正確には「その他の」が表すものの一つに「第 5」がある）.

手汗疱（L30.1 【異汗症[汗疱]】）



腕筋肉痛（M79.1【筋(肉)痛】）

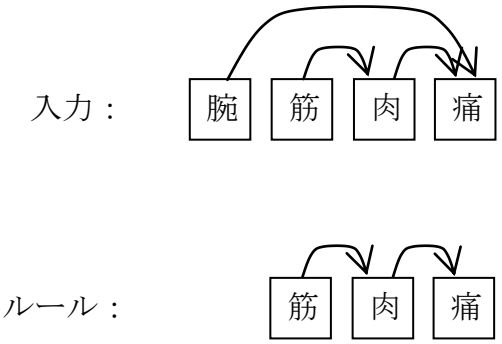


図 16　要因 2（ルールと入力の照合方法の違い）でコーディング可能となった事例

第 5 中手骨頚部骨折（S62.3【その他の中手骨骨折】）

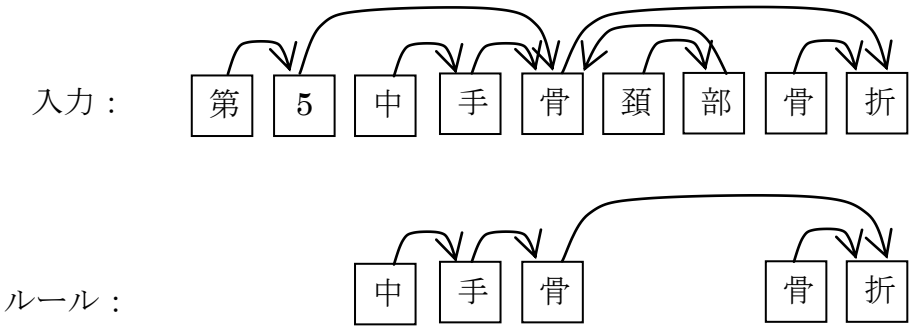
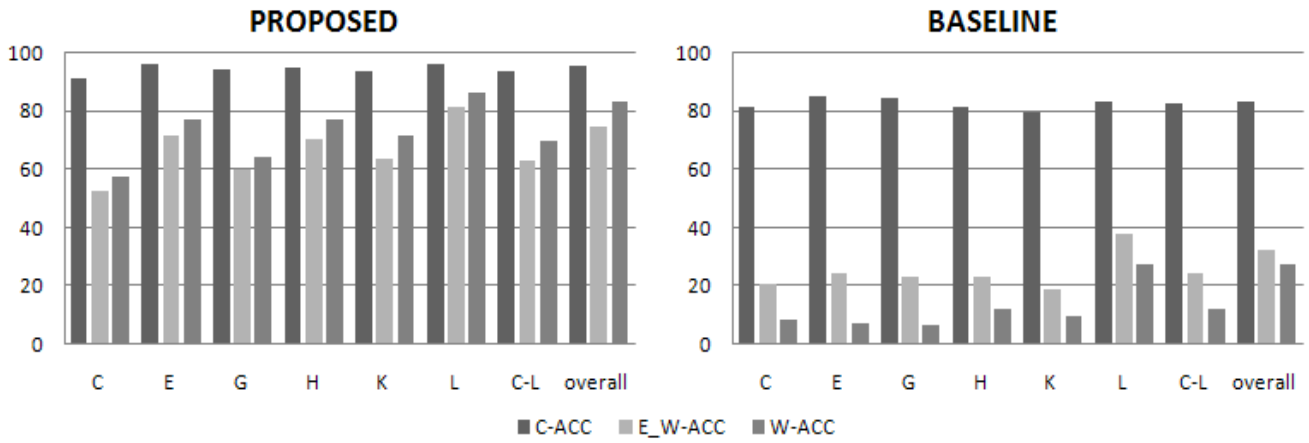


図 17　複合要因でコーディング可能となった事例
入力の中の「第 5」は要因 1,「頚部」は要因 2 による.

4.2. 内部構造解析の精度評価

実験の結果，得られた精度を図 18，解析成功例を図 19，失敗例を図 20 に示す．解析精度は，C-ACC（文字対の解析精度）が 95.4%，W-ACC（語の解析精度）が 83.7%であった． 提案手法は全ての軸について比較手法を上回る精度を達成し，特に W-ACC ではその差は歴然としている．表に示した E_W-ACC は，C-ACC の<平均語長>乗として算出した W-ACC の期待値であり，(1)各係り受け関係はたがいに独立である，(2)文字対の係り受け解析精度は一定である，の 2 点を仮定したものである．提案手法においては全ての軸で W-ACC が E_W-ACC を上回っているが, 比較手法ではその逆となっている. このことは, E_W-ACC でおいた仮定のいずれかもしくは両方が誤っていることを示している.



		C	E	G	H	K	L	C-L	overall
Average Word Length		7.6	9.0	8.8	7.3	7.3	5.3	7.4	6.1
PROPOSED	C-ACC	91.7	96.4	94.4	95.3	94.0	96.3	94.0	95.4
	E_W-ACC	52.8	71.9	60.2	70.4	63.7	81.9	63.3	75.0
	W-ACC	57.5	77.4	64.5	77.0	71.5	86.4	70.1	83.7
BASELINE	C-ACC	81.5	85.5	84.8	81.9	79.7	83.3	82.6	83.3
	E_W-ACC	21.1	24.4	23.4	23.3	19.1	38.0	24.3	32.8
	W-ACC	8.5	7.5	6.8	12.0	9.6	27.6	12.2	27.6

図 18　内部構造解析の精度評価実験結果.

MaltParser による解析を PROPOSED，全ての文字について次の文字に係り受けが存在するとした比較手法による解析を BASELINE で示す.

グラフの縦軸は精度，横軸は実験データの所属クラスである.「C」から「L」は ICD-10 の分類軸，「C-L」は C，E，G，H，K，L の 6 軸を合計したもの，「overall」は「C-L」にそれらの構成要素で語であるもの（例えば「急性 A 型肝炎」に対しては「A 型肝炎」）を追加したデータである.

表の「Average Word Length」は各実験データにおける平均文字列長である.

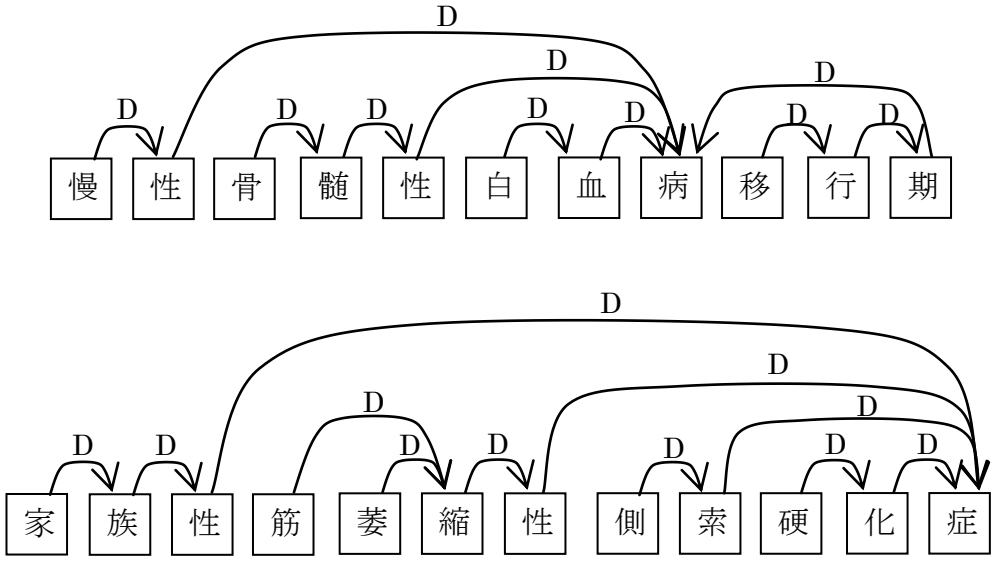


図 19　内部構造解析の成功例.

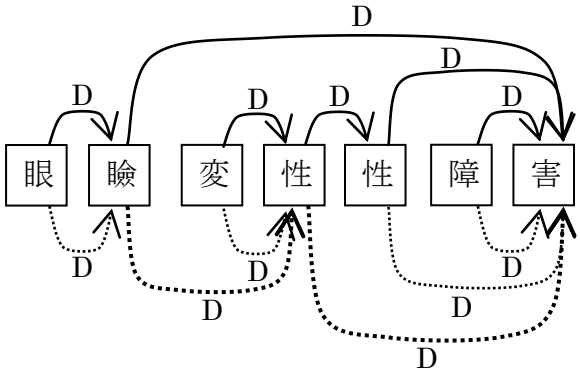


図 20　内部構造解析の失敗例.
実線で正解を，点線で解析結果を示した.「性」と「性」の間に係り受け関係が成り立つと判断するのは非常に難しいと考えられる.

5. 考察

前章で述べた結果から，本研究で提案した疾患名の内部構造表現法が自動 ICD コーディングの精度向上に寄与し，また文字列表現からの自動解析がある程度可能であり実用性が示唆された．一方で，提案手法には限界もあることが分かった．以下その詳細について内容別に論ずる．

5.1. 内部構造表現

5.1.1. 文字を単位語とすることの利点

一般的にテキストの自動処理では形態素（意味を持つ最小単位．一般に複数文字から成る）を単位語とし，文は形態素が重なり合うことなく並べられたものとして表現される．これに対して，本研究では文字を単位語とした疾患名の内部構造表現を提案した．これにより，これまで扱うことのできなかった省略や縮退といった現象を扱うことができるようになった．また，「巨口症」のように形態素列による表現がそぐわない語に対しても，より適切な表現を与えることが可能となった（図 21）．

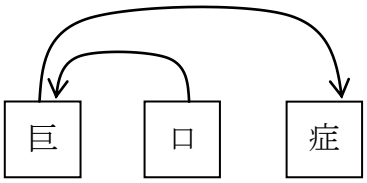


図 21　提案した内部構造表現で適切な表現が可能となった語の例
類似の疾患名として「巨指症」「巨頭症」などが存在するが，もし意味単位で分割すると「ある部位」が「巨大化する症候群」つまり「部位（口，指，頭，など)」+「巨*症」と分割するのが自然である．このような表現は従来の形態素列による表現では不可能であったが，提案手法では自然な表現が可能である．

また，形態素による表現では何を形態素とするかを決めなければならない．しかしこの定義はアプリケーションや個人によっても異なる場合があり，万人に受け入れられる形態素を決めるのは非常な困難を伴う．この問題に対し，提案手法ではテキストの最も小さな構成要素である文字を単位語としたため，形態素定義に依存しない表現が可能であり，また後述するように必要に応じてより長い構成要素へと文字同士をまとめ上げていくことで様々な分割粒度の構成要素を得ることができる（図 22)．従って，提案した内部構造表現はテキスト処理の基盤としての役割も果たすことが期待される．

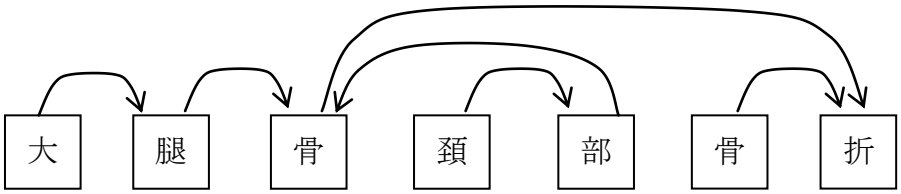


図 22　内部構造表現からの構成要素の切り出し
「腿」に注目すれば「大腿」,「骨」に注目すれば「大腿骨」「大腿骨頚部」を構成要素として切り出すことができる．

5.1.2. 本提案手法の課題

a）表現力の限界

提案した内部構造表現では一般的な係り受け構造と同様に「ある文字について, Modifier はただ一つである」という仮定を置いている．これは複雑な構造を避け自動解析の難易度を抑える効果がある．しかしこの制約は表現力に限界を置くものであり，問題が生じる場合もある．

まず，一部の縮退が 1 つ目の問題である．「術前後 = 術前 + 術後」のように語の最後の文字で縮退が起きている語の表現ができない．これを表現するためには，(1)複数 Modifier を持つことを許す（「術」が「前」と「後」の両方を修飾する），(2)係り受け関係の種類として並列関係を新たに導入する（「1」と「2」が並列関係で結ばれ，「術」が「後」を修飾する），の二つが考えられる（図 23）．いずれにしても問題設定は難しくなるのは避けられず，今後の課題である．

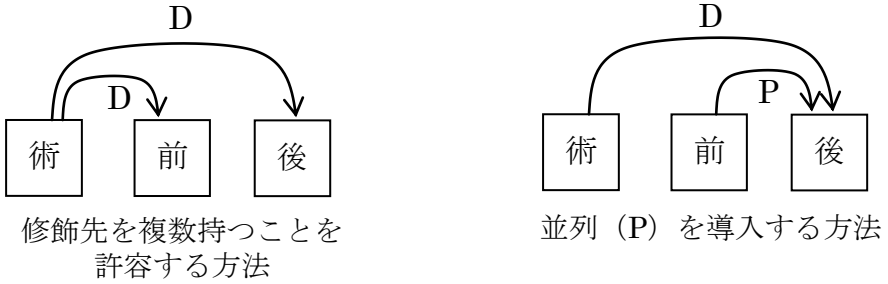


図 23　「術前後」の表現方法

このような縮退現象を含む，より複雑な例として，「第 1・2 腰椎圧迫骨折」が挙げられる．これは「第 1 腰椎」と「第 2 腰椎」における「圧迫骨折」であり，「第 1・2 腰椎」の中で「第」が「1」と「2」の両方を修飾し，更に「腰椎」が縮退している例である．これを提案した内部構造表現に上述の方法をそれぞれ適用すると（図 24），両者はそれぞれ問題を抱えていることが分かる．まず図 24 上部（複数の修飾先を許容）では「第→2」と「1→折」の矢印が交差している．これは一般的に係り受け解析で導入されている非交差制約と呼ばれる制約を破るものであり，制約を外すと解くべき問題の難易度が上がって精度の低下が予想される．一方，図 24 下部（並列 P を導入）では，「腰椎」が縮退していることが表現できていない．これに対して，「並列関係にある複数構成要素が修飾している場合は修飾先が縮退している」というルールを導入することも可能であるが，アドホックな方法であるということ，このルールでは縮退で抜け落ちた文字が「椎」一文字のみではなく「腰椎」であるということが分からないという問題が残る．

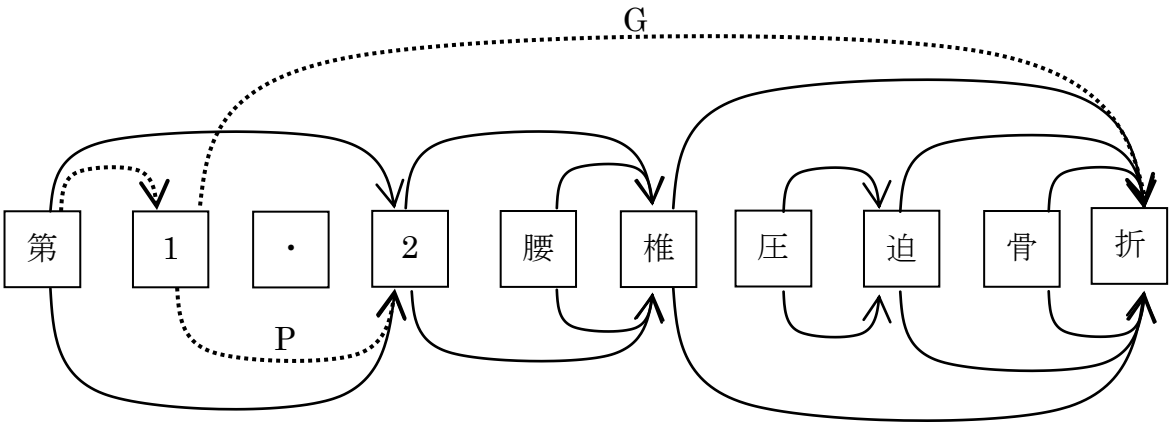


図 24　「第 1・2 腰椎圧迫骨折」の内部構造．
上は修飾先を複数持つことを許容した場合，下が並列 P を導入した場合の表現方法を示す．上下で異なる係り受け関係を点線で示す．「・」はここでは無視している．係り受け関係が D である矢印についてはラベルを略した．

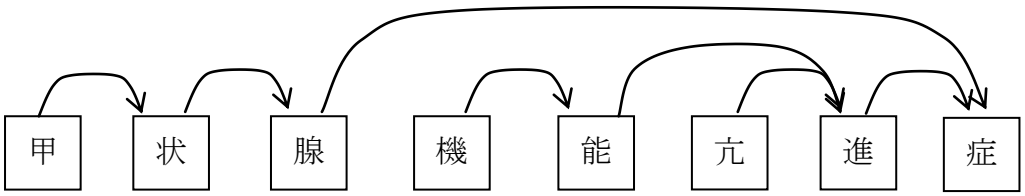


図 25　「甲状腺機能亢進症」の内部構造表現.
内部構造表現をもとにすると,「甲状腺」「機能」という語や「甲状腺（の）症」は取りだせるが,「甲状腺」と「機能」の関係は取りだせない.

「ある文字について，Modifier はただ一つである」という仮定を置いた係り受け表現に起因するもう一つの限界として，実際には存在する構成要素間の意味的な繋がり（係り受け関係）を内部構造表現の中に記述しきれない場合がある．例えば,「甲状腺機能亢進症」の内部構造表現は図 25 である．この中で「腺」に着目すると,「状」からの係り受け関係と「症」への係り受け関係を持っていることから「(甲) 状腺 (の) 症」であることが分かる．一方で「甲状腺機能亢進症」は「甲状腺の機能が亢進する症」であり,「甲状腺」と「機能」は意味的関連を持っている．ところが，内部構造表現では「腺」と「能」の間に係り受け関係が無いために, この意味的関連を内部構造表現のみから読み取ることは不可能である．原因は文字列「甲状腺機能亢進症」の 2 分割が「甲状腺 | 機能亢進症」「甲状腺機能亢進 | 症」のように複数存在するにも関わらず，内部構造表現の付与方法においては前方探索の方針により「甲状腺 | 機能亢進症」という分割を恣意的に選択していることであり，恣意的な選択をしなければならないのは「ある文字について，Modifier はただ一つである」という仮定（制約）を満たすためである．

この問題を解決する方法は 2 つ考えられる．1 つは，上記の制約を排除し，より自由度

の高い文字間関係を記述できるように内部構造表現の定義方法を変更することである．もう１つの方法は，内部構造表現は本研究で提案したものに固定し，「甲状腺機能亢進症」という文脈において「甲状腺」と「機能」の間に関係があることを特定するような上位処理を別途導入することである．両方法は，情報をどの段階で得るか（内部構造解析/上位処理）による違いである．いずれにしても新たな情報を付与しなければならない分，解析精度は低下する可能性が高い．

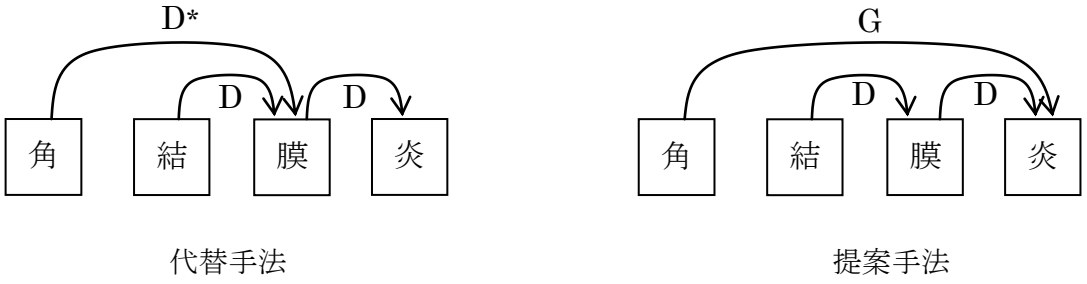
b）機能表現の省略

機能表現とは概念の内容ではなく文や語を構成する機能を持つものであることは既に述べた．例えば「による」「の」「を含む」，あるいは修飾語を作る接尾辞として用いられる「性」（「糖尿病性」など）が挙げられる．特に「性」は「多発性骨折」「多発骨折」のように同じ言葉に対して使われたり使われなかったりする場合がある．ここで，「多発骨折」は「多発性骨折」の「性」が省略されたものと捉えるのが自然であるように考えられる．ただしこの場合，人手で内部構造を定義する際に「これは性が省略された語である」と判断しなければならず，作業者の判断の揺れにより均質なデータの作成が困難となる可能性が高い．一方で，「性」を前処理で削除するという方法も考えられる．しかし「性器」のような機能表現ではない場合があり，「性」がどの意味で使われているのかを判断するステップが必要となってしまう．また同じ接尾辞であっても「糖尿病性腎症」「急性肝炎」のように削除すると不自然である場合もあり，どこまでを削除するべきかという基準を恣意性を排除した形で作成するのは困難であろう．

5.1.3. 発展

a）縮退現象の表現方法

提案した内部構造表現では，縮退を省略と同じ枠組みで扱ったが，図 26 に示すように縮退が起きている場合に重複文字に直接係るよう表現することも可能である．この代替手法は直感に沿っており，また縮退して消えた文字の推定を別途行う必要がないという点で提案手法に対して有利である．しかし一方で，縮退が起きていることが明示できておらず，構成要素の抽出が困難になるという欠点がある．「角→膜」の係り受け関係として縮退を表す新たなラベルを導入すればこの問題は回避できるが，ラベルの種類の増加は解析精度を低下させるものである．学習データが十分量あれば影響は少ないが，現状では十分な学習データの用意は難しいため，本研究ではラベルを増やすことを避けた表現方法を採用した．



代替手法　　提案手法

図 26　縮退現象の表現方法の代替案

b）構成要素のいずれも Head となりうる場合

例として「糖尿病性腎症」を挙げる．内部構造は図 27 のようになる．

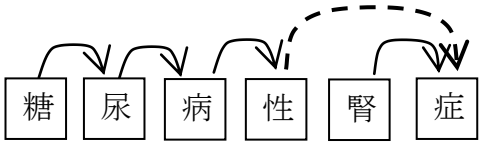


図 27　「糖尿病性腎症」の内部構造

ここで,「糖尿病」と「腎症」のいずれが Head であるかということが疑問に上がる．本研究で採用した Head の決め方に則れば「糖尿病性腎症」は「is-a 糖尿病」「is-a 腎症」のいずれも満たす．これを忠実に表現するには，両者の係り受け関係として「並列」を導入する必要がある．しかし一連のテキスト解析の中で内部構造表現が扱うべき範囲は「詳細な医学知識無しで，表層文字列のみから取り出せる情報を明示すること」である．ここでは表層文字列に「性」の文字があるため,「糖尿病」に修飾語としての役割が付与されていると捉えるのが妥当であろう．従って「is-a 糖尿病」が成立するか否かの判断は内部構造表現の扱う対象外であり，本研究は「腎症」を Head と捉えるのが正しいという立場を取る．

c) Head と Modifier の基準

本研究では Head を決める際，part-of 関係よりも is-a 関係を優先した．これは，例えば「大腿骨」の Head として「大腿」よりも「骨」のほうが直観に沿ったものであるからである．しかし「大腿の骨折」という表現は不自然でない表現であるが，これは「大腿骨の骨折」で縮退もしくは省略が起きたのではなく,「大腿の（一部である骨における）骨折」

と捉えるほうが自然であろう．上位下位関係と部分全体関係のどちらを優先させるべきかについての議論が今後必要である．

5.2. 自動 ICD コーディング

既存の自動 ICD コーディング手法と提案手法を組み合わせることで，コーディング精度が少なくとも 3%向上した．すなわち，提案手法により既存の自動 ICD コーディング手法では解決できなかった例の一部を解くことができた．既存手法に対して内部構造表現が精度向上に寄与する例は全て，解剖部位の部分全体関係が内部構造表現から判断できる場合であった．すなわち，予め外部知識としてあらゆる語の部分全体関係を記述しておかなくても解析可能となる場合が存在するということである．解剖部位は「大腿骨+頚部」のように修飾語を加えていくことが可能であるため，多様な表現が存在しうる．全ての可能性を網羅した外部知識を用意するのは多大な労力を必要とするものであり，そのような外部知識を必要としない本手法が貢献する部分である．

また ICD コーディングという応用において省略・縮退を扱うには，(1) コーディング用外部知識として省略・縮退がすでに起きた状態の語をオントロジに収載する，(2) 言語処理の段階で自動的に復元する（提案手法の範疇），という二つの手段が考えられる．(1)の場合，例えば標準病名と構造化 ICD に現れるラベルの差分を省略・縮退した文字として取り出し，それを新たなラベルとして構造化 ICD に追加するという方法が考えられる．この方法は確実にコーディング精度を向上させるものであるが，一方で当該 ICD コードのみの

情報を拡充しているに過ぎず，他 ICD コードに対応する疾患名で同じあるいは類似の現象が起きていてもこれを正しく扱うことができない．一方(2)の場合は，100%の精度ではないにしろ，他 ICD コードであっても省略・縮退を展開することができる．2 つの手法を組み合わせることで，自動 ICD コーディング精度は向上が可能となるであろう．

既存手法は修飾語の順序のみが異なる表記を同一のものと見なすため，修飾語の順序の違いを吸収するものであるが，一方でこれらが異なる概念であっても同一と見なしてしまうという欠点がある．例えば「嚢胞腎」と「腎嚢胞」の 2 つの異なる概念を共に「嚢胞，腎」という 2 語の集合として扱う．2 語を別の語として認識するためには，前者を「嚢胞→腎」，後者を「腎→嚢胞」と係り受け関係を考慮に入れなければならない．このような問題は，本研究で提案したような内部構造を用いなければ解決できない．（ただし，ICD コーディングに限って言えば，必ずしも全てを自動で行う必要はなく，候補となる ICD コードを作業者に提示するという立場を取ることも可能であり，この場合は「嚢胞腎」と「腎嚢胞」を区別できないことはそれほど大きな問題とはならない．）

5.3. 内部構造解析

解析器の学習用に作成したデータの量は少ないものであったにも関わらず，文字対の係り受け関係に対して 95%と高い精度での解析が可能であった．実際に内部構造情報を利用しようとした場合には，文字列で表現された医学用語を自動で内部構造表現に変換しなければならないが，実験の結果，解析は精度よく行えることがわかったので，内部構造表現

の実用性が示されたことになる．なお，一般の文に対する係り受け解析器が，本研究では使用できない品詞情報という強力な素性を使用した上で文節間の係り受け関係（本研究における文字対の係り受け関係に相当する）9 割程度の解析精度であることを考えると，本研究における解析精度は非常に良い成績であるといえる．

一層の精度向上には，素性とアルゴリズムの工夫，学習データ量の増加が寄与するであろう．提案手法において W-ACC の値が E_W-ACC より高かったことから，解析対象となる文字対によって解析精度が異なると考えられる．精度向上のためには，解析精度の低い文字対がどのような特徴を持つのかを調査し，その結果を素性やアルゴリズムに反映させる必要がある．手法としては，本研究では決定的な手法，すなわち語全体を見渡すことなく語の先頭から順番に係り受け関係を決めていく方法を取ったが，この方法では全体としての最適化はなされない．例えば「肺炎症性偽腫瘍」という語に対して語頭から解析を行った場合，解析器はまず「肺」と「炎」の間に係り受け関係があるか否かを判断する．この時，素性として使った辞書情報に「肺炎」が掲載されていれば，解析器は「肺→炎」という係り受けがあると判断すると予想される．しかしこの語は「肺における炎症性偽腫瘍」であるため「肺→瘍」が正解である．これを正しく解くためには，係り受け関係の有無・種類を判断する際に，単位語全てを見比べ，相対的に適切な単位語を係り受け先として選ばなければならない．このような特徴を持つアルゴリズムとして Maximum Spanning Tree[37]や Cascaded Chunking[7]が提案されている．

実験においては標準病名マスターに収載されている疾患名を材料として使用した．最初の対象として最も標準的な材料を選択するのは妥当であるが，現実に存在する疾患名は標

準病名に比べ表現が多様であるはずである．今後，標準病名マスターに収載されていない疾患名を材料とした解析器の学習や，解析アルゴリズムの改良によって対応していく必要がある．

また，提案手法の最も大きな特長である省略・縮退現象については，用意したデータでの出現回数が少なく，これに対しても同等の精度で解析が可能であるかどうかは不明である．今後，省略・縮退現象を精度よく扱う改良が必要である．

5.4. 内部構造表現の応用

ICD コーディング以外のアプリケーション

提案した内部構造表現はアプリケーションに依存しない汎用的なものであり，本研究で述べた自動 ICD コーディングのみならず，DPC や MedDRA 等，他の分類体系へのコーディングにおいても有用であると考えられる．また応用先はコーディングのみに限定されるものでもない．

内部構造表現はその語が示す概念を表現するものではないものの文字列表現より情報量が多く，解析器を作ってしまえば文字列表現を解析することで人手を介さず利用できるというのが大きな利点である．ただし解析手法を工夫しても精度にはある程度の限界があると考えられるため，ある程度の誤りを許容できる場面での使用が望ましい．

一例として，医師が入力した疾患名のチェックが挙げられる．本研究で解析器の学習データとして標準病名を使用しており，解析器は入力に対して標準病名に現れるような構造を付与しようとする．この時，解析器が解析不可能という結果を出すことで，医師の入力

する疾患名を「標準病名らしく」揃えることができる．本稿で述べた解析器では解析不可能という出力はできないが，今後これに対応していくようなアルゴリズムの改変を考えている．

他の例として，大量の文書に対しての検索や統計のための用語抽出の際に有用となる可能性がある．例えば「A 型肝炎」を検索した時，文字列一致では「A 型急性肝炎」は見つからないが，内部構造を考慮した検索を行えば発見することができる．

また，医療への患者参加やパターナリズムからの脱却が社会的に認識され，また情報の流通が盛んである現代においては，非医療従事者が医学情報に触れる機会は多い．しかし知識を持たない非医療従事者にとっては医学知識へのアクセスや情報の理解は困難を伴うものである．このような状況で，文字列から一歩意味へ踏み込んだ内部構造情報が寄与する場面があると考えられる．今後この可能性について検討する必要がある．

表意文字，他言語

提案した内部構造表現は文字を単位語としているため，原理的には言語を選ばない表現方法であると言える．ただし一般にアルファベットや平仮名などの表音文字は表意文字に比べて数が少ないため，単位語の持つ情報が少なく，自動解析の難易度が高くなるという難点がある．日本語医学用語の場合は漢字の使用が非常に多いため，文字を単位語とすることが比較的自然であると言える．

表意文字である漢字を主とする中国語では文字を単位語とした解析が日本語よりも自然であると考えられる．実際に，文解析では従来の形態素解析と形態素間の係り受け解析と

いう枠組みだけでなく，本研究での提案と同様に文字を単位とした係り受け解析を試みているものもある[41,42].

内部構造表現を基盤としたテキスト処理

提案した内部構造表現は，言葉の持つ意味の表現というよりも，言葉を構成する文字の間の関係を記述するものである. 内部構造表現で導入した文字の間の関係（係り受け関係）は，省略・縮退が起こっているか否かの区別がなされているだけであるが，実際には語を構成する関係と 2 語を結びつける関係が混在している．例えば「びらん」と「食道癌」は，文字内容を無視すれば内部構造は等しい（図 28）．しかし形態素列として捉えた時には「びらん」は 1 形態素，「食道癌」は「食道」と「癌」の 2 形態素と考えられるから，「ら→ん」の係り受け関係と「道→癌」の係り受け関係は本質的に異なるものである．内部構造表現のみを与えられた場合，前述のように任意の粒度で構成要素を切り出すことができるが，これが語である保障はない．びらんと食道癌の例でいえば，中央の文字に着目すると「びら」と「食道」が得られ，前者は語ではなく後者は語である．

図 28　提案した内部構造表現では同じ構造となるが, 意味的には構造に違いがある例

本研究で提案した自動 ICD コーディング手法のように，内部構造表現を照合する対象が語の情報を持っている場合，内部構造表現から取りだされる文字列の中で語とならないものは照合の際に除外されて害を成すことはないが，照合先が存在しないような応用においては問題となる．単位語として意味を持つ文字列である形態素を用いた形態素解析では 1) 形態素定義の困難, 2) 省略・縮退を考慮しない, という問題が発生することは既に述べた．形態素解析の問題を解決するための内部構造表現であり，形態素解析は代替手段にはならない．これらの問題の解決策としては，内部構造表現を対象とした形態素解析，すなわち文字をノードとするグラフの分割が挙げられる．あるいは内部構造表現内の係り受け関係に強度を付与し，強度の強い順に文字を結びつけていくことで，段階的な形態素解析を行うことも可能であろう（図 29）．さらに，内部構造表現は従来の形態素間の係り受け関係を内包しているので従来の係り受け解析が必要ない．

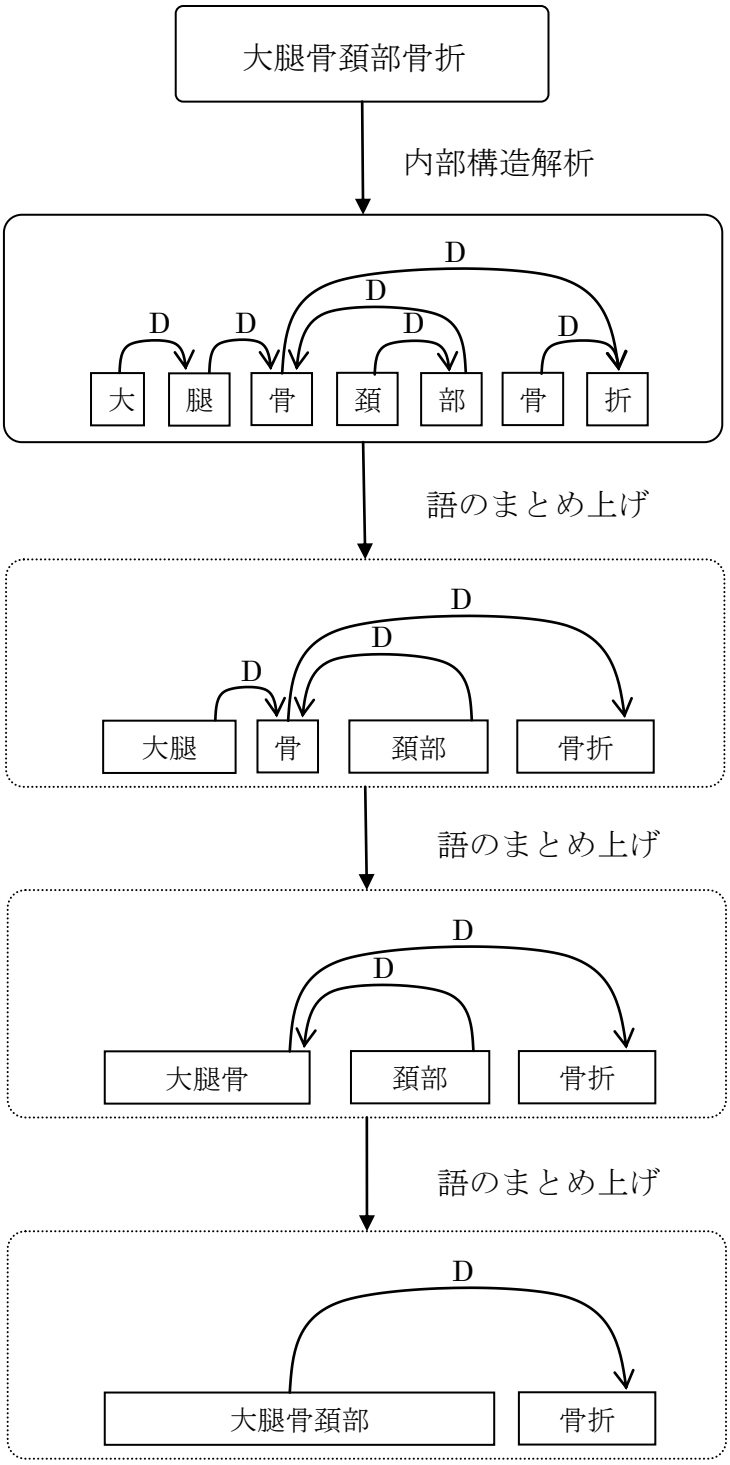


図 29-1. 内部構造表現を基盤としたテキスト処理の案.
文字列として入力された語を内部構造解析により内部構造表現へ変換する. 内部構造表現に現れる係り受け関係に従って語をまとめ上げていくと, さまざまな粒度の形態素とその係り受け関係が同時に得られる.

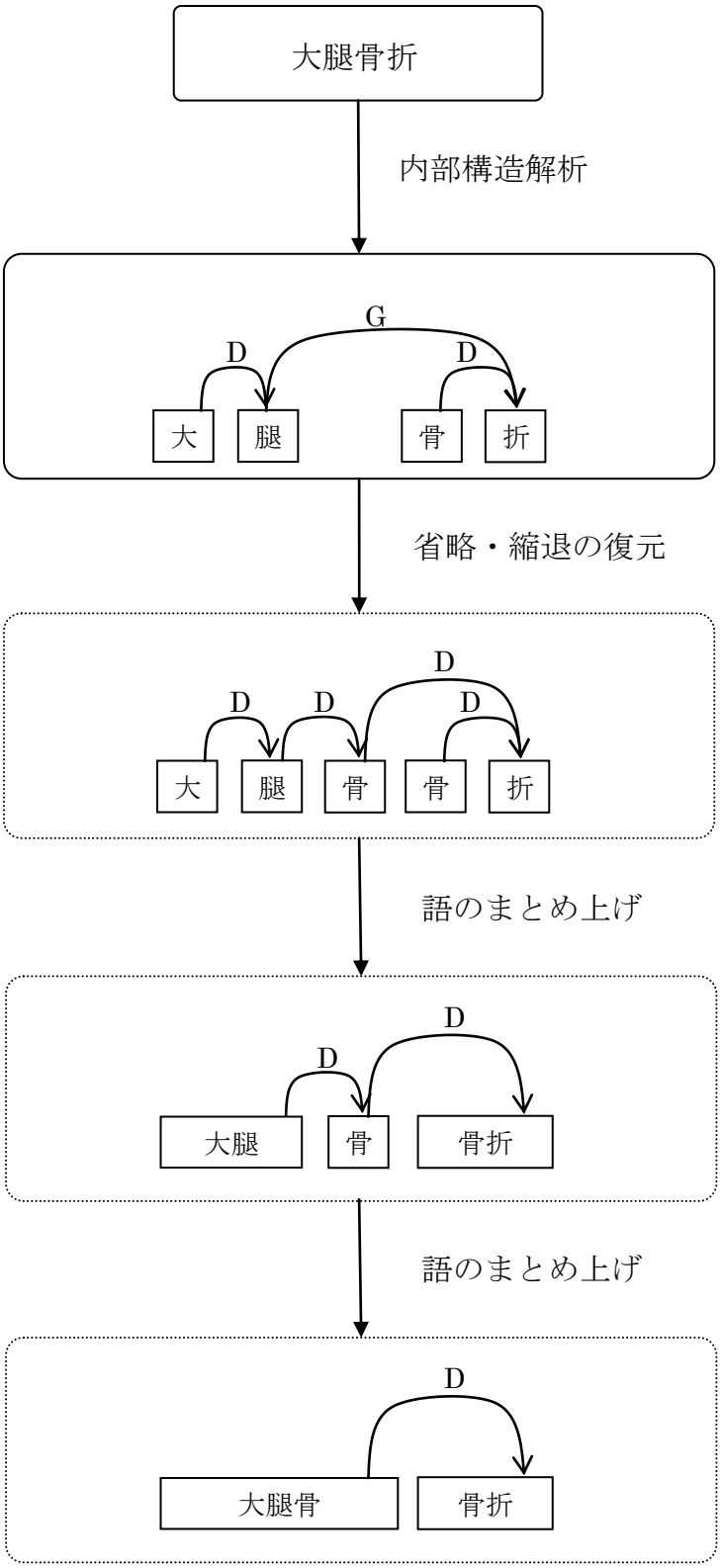


図 29-2. 内部構造表現を基盤としたテキスト処理の案.
省略・縮退がある場合（内部構造表現でラベル G が出現した場合）は語のまとめ上げ処理の前に復元処理を行う.

6. 結語

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を提案し，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．

謝辞

本研究を進めるにあたっては，多くの方のご助力をいただきました.

まず，一年の間お世話になりました奈良先端科学技術大学院大学 自然言語処理学講座 松本 裕治教授，ならびに同講座の皆さまには，複合語の内部構造や係り受け解析アルゴリズムについてご指導いただき，粘り強くご議論をいただきましたことを感謝いたします.

また，医療情報学と自然言語処理学のいずれにも造詣が深い 知の構造化センター 荒牧 英治先生には，日々細やか且つ明るいご指導・ご鞭撻をいただきましたことを感謝いたします．同じく両分野について造詣が深く，ICD コーディング実験に関してデータを提供いただきました医学系研究科 今井 健先生には，真摯なご指摘・ご指導をいただきましたことを感謝いたします.

そして，ご指導いただきました医療情報経済学分野 大江 和彦教授に心より感謝の意を表します．狭くなりがちな視野に対して，常に研究の方向性を示していただいたことで，医療情報学の研究としてこのような形にまとめることが出来ました.

最後に，医療情報経済学教室の皆さま，また支えとなっていただいた方へ感謝の意を表して，謝辞といたします.

参考文献

[1] JUMAN. http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html (2010.1.6.参照)

[2] ChaSen. http://chasen-legacy.sourceforge.jp/ (2010.1.6.参照)

[3] Masayuki Asahara, Yuji Matsumoto : Extended Models and Tools for High-performance Part-of-Speech Tagger. Proceedings of the 18th conference on Computational linguistics, Volume 1, pp.21-27, 2000.

[4] KNP. http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html (2010.1.6.参照)

[5] Sadao Kurohashi and Makoto Nagao. KN Parser: Japanese Dependency/Case Structure Analyzer. Proceedings of The International Workshop on Sharable Natural Language Resources, pp.48-55, 1994.

[6] CaboCha. http://chasen.org/~taku/software/cabocha/ (2010.1.6.参照)

[7] Taku Kudo and Yuji Matsumoto. Japanese dependency analysis using cascaded chunking. Proceedings of the Sixth Conference on Computational Language Learning (CoNLL), Volume 20, pp.1-7, 2002.

[8] Taku Kudo, Yuji Matsumoto. Fast Methods for Kernel-Based Text Analysis. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Volume 1, pp.24-31, 2003

[9] WordNet. http://wordnet.princeton.edu/ (2010.1.6.参照)

[10] Medical Subject Headings: MeSH. http://www.nlm.nih.gov/mesh/meshhome.html (2010.1.6.参照)

[11] Systematized Nomenclature of Medicine-Clinical Terms: SNOMED-CT. http://www.ihtsdo.org/snomed-ct/ (2010.1.6.参照)

[12] Foundational Model of Anatomy: FMA. http://sig.biostr.washington.edu/projects/fm/ (2010.1.6.参照)

[13] Rogers JE, Roberts A, Solomon WD, van der Haring E, Wroe CJ, Anstra PE and Rector AL. GALEN Ten Years On: Tasks and Supporting tools. Proceedings of MEDINFO2001, pp.256-260, 2001.

[14] Unified Medical Language System: UMLS. http://www.nlm.nih.gov/research/umls/ (2010.1.6.参照)

[15] Werner Ceusters, Barry Smith, Anand Kumar, Christoffel Dhaen: Ontology-Based Error Detection in SNOMED-CT®, Proceedings of MEDINFO2004, 482-486, 2004.

[16] 梅木定博, 後藤智範. 辞書見出し語の 7 文字漢字熟語を対象とした語基構成の解析. 情報処理学会研究報告 自然言語処理 研究報告 No.184, pp.113-118, 2008.

[17] 小山照夫, 大江和彦. 医学専門用語の構造解析. 学術情報センター紀要, No.6, pp.115-124, 1994.

[18] 小林義行, 徳永健伸, 田中穂積. 名詞間の意味的共起情報を用いた複合名詞の解析. 自然言語処理, Vol.3. No.1, pp.29-43, 1996.

[19] 韓東力, 伊藤毅志, 古郡廷治. 要素間の依存関係に基づく複合語の構造分析. 電子情報通信学会論文誌 D Vol.J86-D2, No.5, pp.706-714 , 2003.

[20] 富樫秀夫, 栗原勝, 折井孝男. 医薬品添付文書情報の解析. 医療情報学 26(2),

pp.129-134, 2006.

[21] 竹内孔一，乾健太郎，藤田篤，竹内奈央，阿部修也:分類の根拠を明示した 動詞語彙概念構造の構築，自然言語処理研究会 2005-NL-169, 2005.

[22] 工藤拓. 形態素周辺確率を用いた分かち書きの一般化とその応用. 言語処理学会年次大会発表論文集 11, pp.592-595, 2005.

[23] International Classification of Diseases (ICD). http://www.who.int/classifications/icd/en/ (2010.1.6.参照)

[24] 標準病名マスター. http://www.dis.h.u-tokyo.ac.jp/byomei (2010.1.6.参照)

[25] Eiji Aramaki, Takeshi Imai, Kengo Miyo, Kazuhiko Ohe. Orthographic Disambiguation Incorporating Transliterated Probability. International Joint Conference on Natural Language Processing, pp.48-55, 2008.

[26] Kevin Knight and Jonathan Graehl. Machine transliteration. Computational Linguistics, 24(4), pp.599-612, 1998.

[27] Eiji Aramaki, Takeshi Imai, Masayuki Kajino, Kengo Miyo, Kazuhiko Ohe. A Statistical Selector of the Best among Multiple ICD-coding Methods. Proceedings of MEDINFO2007, pp.645-649, 2007.

[28] Tagliabue G, Maghini A, Fabiano S, Tittarelli A, Frassoldi E, Costa E, Nobile S, Codazzi T, Crosignani P, Tessandori R, Contiero P. Consistency and accuracy of diagnostic cancer codes generated by automated registration : comparison with manual registration. Popul Health Metr, Vol.4, pp.10, 2006.

[29] Pakhomov SV, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. J Am Med Inform Assoc, Vol.13(5): pp.516-25, 2006.

[30] 今井健，荒牧英治，梶野正幸，美代賢吾，大江和彦．臨床医学分野における用語概念間の関係情報を用いた自動 ICD コーディングに関する研究．人工知能学会全国大会論文集, 22, 2E3-03, 2008.

[31] He'ja G, Surja'n G, Luka'csy G, Pallinger P, Gergely M. GALEN based formal representation of ICD10. Int J Med Inform, Vol.76(2-3), pp.118-23, 2007.

[32] Fabry P, Baud R, Ruch P, Le Beux P, Lovis C. A frame-based representation of ICD-10. Stud Health Technol Inform, Vol.95, pp.433-8, 2003.

[33] Nivre, J., J. Hall and J. Nilsson. Memory-Based Dependency Parsing. In Ng, H. T. and Riloff, E. (eds.) Proceedings of the Eighth Conference on Computational Natural Language Learning, pp. 49-56, 2004.

[34] Michael A. Covington. A fundamental algorithm for dependency parsing. Proceedings of the 39th Annual ACM Southeast Conference, pp.95-102, 1991.

[35] R. McDonald, F. Pereira, K. Ribarov, and J. Hajiˇc. Non-projective Dependency Parsing using Spanning Tree Algorithms. Proceedings of HLT/EMNLP 2005.

[36] Masakazu Iwatate, Masayuki Asahara, Yuji Matsumoto. Japanese dependency parsing using a tournament model. Proceedings of the 22nd International Conference on Computational Linguistics, Volume 1, pp. 361-368, 2008.

[37] Fei Sha, Fernando Pereira. Shallow Parsing with Conditional Random Fields. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Volume 1, pp.134-141, 2003.

[38] MaltParser. http://maltparser.org/ (2010.1.6.参照)

[39] Nivre, J., J. Hall and J. Nilsson. MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. Proceedings of the fifth international conference on Language Resources and Evaluation, pp.2216-2219, 2006.

[40] 今井健, 荒牧英治, 梶野正幸, 美代賢吾, 大江和彦. 階層分類情報を用いた疾患オントロジーの半自動構築, 医療情報学, Vol.27 Suppl., pp.700-3, 2007.

[41] Hai Zhao. Character-Level Dependencies in Chinese: Usefulness and Learning. The 12th Conference of the European Chapter of the Association for Computational Linguistics, pp.879-887, 2009.

[42] Hwee Tou Ng and Jin Kiat Low. Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? Proceedings of EMNLP 2004.

図表一覧