専門用語の内部構造解析

山田恵美子 a)b)　　松本裕治 a)
a) 奈良先端科学技術大学院大学　情報科学研究科
b) 東京大学大学院　医学系研究科
email: emiko-tky@umin.net, matsu@is.naist.jp

1　はじめに

特定分野の文章において専門用語は特有の意味を保持しており、それを知ることはその文章を解析する際に有用である。専門用語は複合名詞であるものが多く、したがって図１のような内部構造を持っている。内部構造とは語の構成要素とその結合の順序であり、これを知ることは意味を推測するうえで役に立つ。しかし専門用語は数多く存在しており、また動的に作られうるものであるため、予め全ての語に内部構造を記述しておくのは現実的ではない。本研究の目的は、専門用語の内部構造で見られる複雑な現象を調べること、また内部構造のタグ付けを行うための方法を提案することである。対象としたのは疾患名を中心とした生命科学分野の用語である。本稿では後で説明するように用語の内部構造を係り受け構造によって表現するが、一般の文で使われる係り受け構造だけでは表現しきれない構造として、後ろから前への係り受けや、「大腿骨折＝大腿骨＋骨折」「角結膜＝角膜＋結膜」のような縮退が起こる場合がある。このような構造を表現するのに必要な表現方法を提案し、これを文字単位の係り受けと捉えることが可能であることを示す。また、SVM による自動解析の試みを報告する。

2　専門用語

本稿では生命科学分野の専門用語、特に疾患名や解剖部位の用語の内部構造に主眼を置く。ここで内部構造とはその語を構成する構成要素の間の係り受け関係とする（図 2）。

専門用語が一般の複合名詞と異なる点として、以下のようなことが挙げられる。まず専門用語とは特定の領域内で使われるものであり、その分野の文化に依存して独自の内部構造を持ちうる。例えば「糖原病 I 型」では「I 型」が「糖原病」に係っているし、「角結膜炎」の「角結膜」とは「角膜」と「結膜」が並列に結びついたうえで文字の

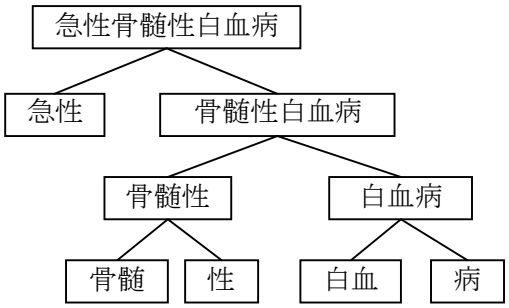


図 1　複合語の内部構造

縮退が起こっているものである。このような構造は、従来の形態素への分割と係り受け構造をそのまま適用しても表現しきれない。

また、ある程度複雑な概念を限られた文字数で表すため、構成性が弱いと考えられる。従って内部構造を一意に決定するのが困難である場合がある。例えば「全前脳症」の「全」はどこに係るだろうか。「前脳」とはヒトの発生段階で脳の一部として存在し、複数の組織に分化する部位である。「全前脳症」とはこれが分化せずそのまま残ってしまったために奇形が生じるという疾患である。この場合、「前脳が全部そのまま残っている」ということから「全」は「前脳」に係るのが正解となる。

複合名詞の内部構造に関する研究として、漢字熟語内の品詞列・係り受けの調査[1]、医学用語を対象とした用語構造解析[4]、意味クラスの共起情報を用いた構造解析[3]、相互情報量を用いた構造解析[5]などが挙げられる。しかしいずれも従来の形態素解析・係り受け解析の域を出ておらず、本研究で扱う専門用語には不十分である。

3　内部構造の表現方法

3.1　係り受け木

既に述べたように複合語の内部構造とは構成要素とその間の関係から成る。分割された構成要

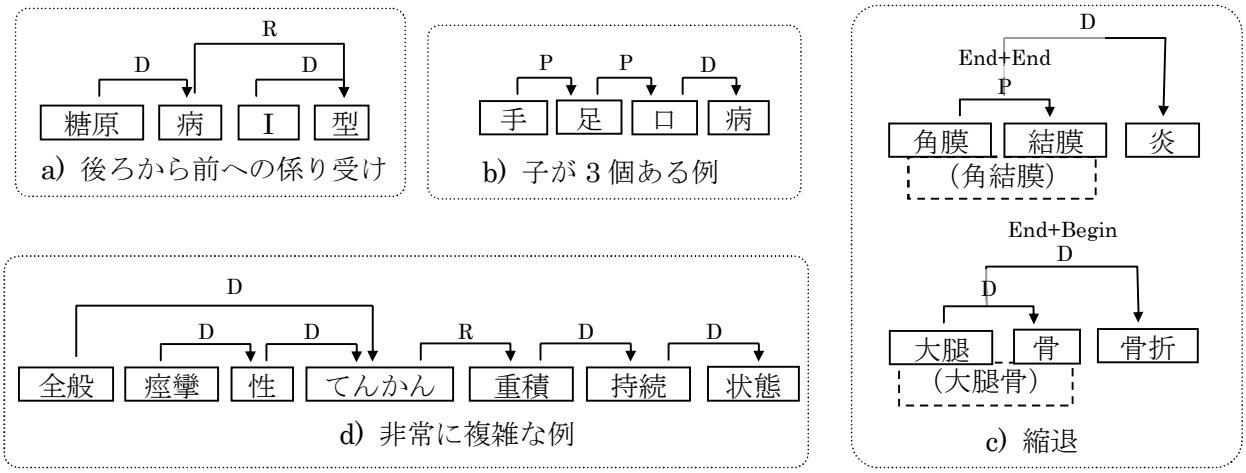


図 2　ラベル付き係り受けによる内部構造表現

素間の関係を以下の 4 種類に分類した。

D：前から後ろへの係り受け
　　例：急性 => 肺炎（急性肺炎）
R：後ろから前への係り受け
　　例：糖原病 <= I 型（糖原病 I 型）
P：並列
　　例：脊髄 + 小脳（脊髄小脳変性症）
U：上記以外の結びつき
　　例：B + 1 + 6（B16 メラノーマ細胞）

図 1 のような内部構造は、枝に上記 4 種のラベルのいずれかを付与した係り受け木で表現できる（図 2)。日本語の係り受け解析では一般に「自身よりも後の形態素に係る」「1 つの形態素が複数の係り先を持たない」「交差が起きない」の 3 つの制約を利用する。1 つ目の「自身よりも後の形態素に係る」に関して、R（後ろから前への係り受け）は従わないことになるが、図 2a のように係り受けの方向は前から後ろに固定し、ラベルで係り受けの種類を表現すればこれは解消できる。2 つ目の「1 つの形態素が複数の係り先を持たない」を有効にするため、1 つの構成要素に対して R で係る構成要素は 1 つであると仮定した。

また、構文木は原則 2 分木であるが、ラベルが P または U である時には子ノードは 2 個以上ある場合がある。この場合は同じ係り受け関係の連続によって表現する。上記の「B+1+6」のほか、図 2b「手足口病」の「手+足+口」のような 3 語が並列関係にある場合がそれにあたる。

上記の表現方法は、今回対象としたデータに対して十分な表現力を持っていた。ラベルを導入したことによって、内部構造解析はラベル付きの係り受け解析として捉えられる。

3.2　文字単位の係り受けによる表現

ここでは 3.1 で触れなかった構成要素への分割について述べる。語を分割する際には文字の縮退を考慮する必要がある。ここで縮退とは「大腿骨折（=大腿骨＋骨折)」「角結膜（=角膜＋結膜)」のように、複数の構成要素が結合する際、オーバーラップしている部分が 1 つに纏められる現象を指す。図 2 に示すとおり、縮退についてもラベルを付与することとした。「End+End」は構成要素の最後の文字が縮退していることを意味し、

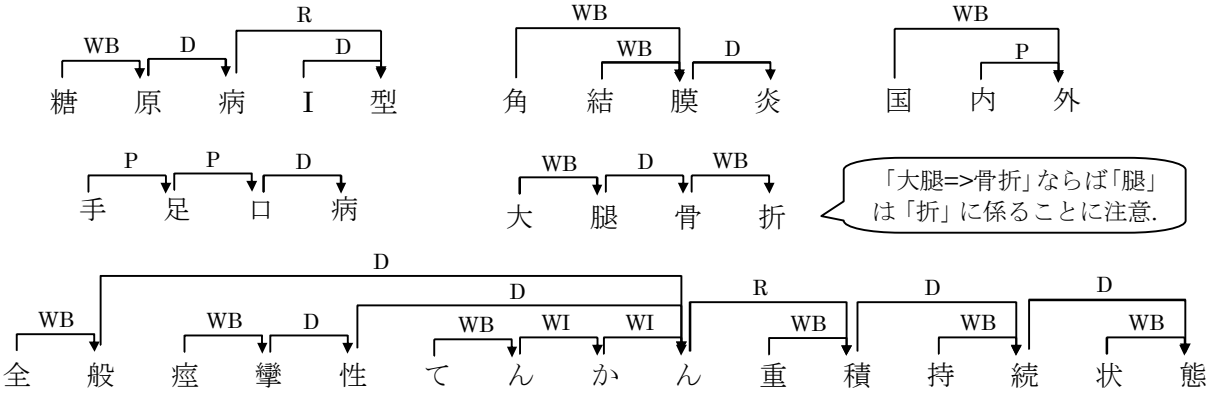
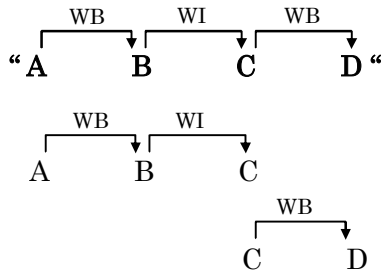


図 3　文字単位のラベル付き係り受けによる内部構造表現

(a) “C”が縮退する場合（ABC+CD）



(b) “BC”が縮退する場合（ABC+BCD）

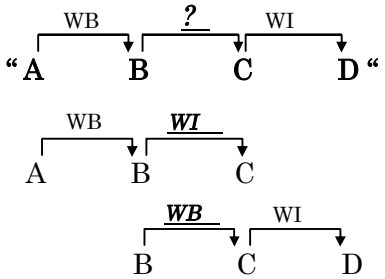


図 4　縮退が起きた時の表現

「End+Begin」は構成要素の 1 番目の最後と 2 番目の最初の文字が縮退していることを意味する。このほかに「国内外（=国内+国外）」のように最初の文字で縮退が起きている場合にはラベルとして「Begin+Begin」を付与する。縮退が起こっていない時にはこのラベルは付与されない。

縮退を自然に表現するために、文字単位の係り受けによって内部構造を表現した。ここでは 3.1 で述べた 4 種の係り受けの他、構成要素そのものを作る係り受けを二種類（WB、WI）追加する。WB は構成要素の先頭部分、WI はそれ以外の部分での係り受けである。図 3 に文字単位の係り受けによる内部構造表現を示す。図中で縮退が起きている 2 語について、「角結膜炎」では「角膜」と「結膜」が P で結ばれることが、「大腿骨折」では「大腿骨」が「骨折」に D で係ることが明示されていないが、縮退が起きている時には以下のようにラベルを決める。前者のような構成要素の同じ部分の文字が縮退する時には P、後者のように一つ目の構成要素の最後と二つ目の構成要素の最初の文字が縮退する時には D とする。

また、この表現方法では 1 回の縮退で消える文字は必ず 1 文字でなければならない。仮に 2 文字が縮退すると、縮退する 2 文字の間に 2 つの関係が同時に成立し、これを表現することができない（図 4b）。現実には 2 文字以上の縮退は起こらないと思われるため、この制約が問題となることはない。

「国内外（=国内+国外）」のような並列かつ最初の文字が縮退している場合は「国内＋国外」としてしまうと「国」の係り先が 2 つになってしまうため、「国 => (内 + 外)」（先に「内外」を並列関係で結び、そこへ「国」が係る）というように表現することとした。

この表現方法を用いると文字の縮退を自然に表現することができる。また従来の形態素解析と係り受け解析に相当する処理を一つの処理に纏めることができる。

4　実験

前節で述べたとおり、文字の間の係り受け関係は 6 種類あるが、今回はその種類を考慮せず、2 つの文字の間に係り受け関係が存在するか否かを判別するタスクとして実験を行った。

データとしてライフサイエンス辞書（以下 LSD）2008 年 4 月版を利用した。LSD は生命科学分野の専門用語辞書で、2008 年 4 月版では日本語 94707 語、英語 83956 語を掲載している。掲載語の一部には文献管理用に開発されたシソーラスである MeSH（Medical Subject Headings）のコードが付与されている。このうち疾患を表すコードの付与された八文字以上の日本語から 251 語に対してタグ付けを行った。この時、タグ付けした語の構成要素の内部構造も同時に保存される。例えば「急性骨髄性白血病」のタグ付け作業は図 1 の一番上から順に行われるが、この時 LSD に掲載されている「骨髄性白血病」「骨髄性」「白血病」も同時に内部構造が定義される。従って、タグ付けされた語は疾患名の他に解剖部位名や物質名を含む。これらを合わせるとタグ付けされた語は合計 794 語である。

このデータを用い、文字対に対して係り受け関係の有無を識別するために TinySVM を用いて学習を行った。カーネルは線形、使用した素性は表 1 のとおりである。

次にこの識別器を用い、係り先が一つ、交差は起こらないという制約を入れ、前方から順に係り先を決定する実験を行った。係り先は SVM の出力数値が最も大きなものを選択するようにした。

ライフサイエンス辞書プロジェクト
http://lsd.pharm.kyoto-u.ac.jp/index.html

表 1 係り受け判定に用いた素性

文字情報	係り元/係り先候補の漢字
	係り元/係り先候補の文字種（ひらがな、カタカナ、漢字)
	係り先候補の一文字後がカタカナかどうか
文脈情報	係り元と係り先候補の文字の距離
	元の文字列中での位置（先頭、途中、最後）
辞書情報	係り元の文字と係り先候補の二文字から成る語が辞書に掲載されているかどうか
	係り元/係り先候補で終わる語で、かつ元の文字列中に含まれる語が辞書に掲載されているかどうか

表 2 係り受けの判定精度

Data	文字対	語
Baseline	86.9%	28.6%
System	93.7%	55.6%

表 2 に結果を示す。ベースラインとして全ての文字が次の文字に係るとしたときの値を示している。なお、システムの値は 10 分割交差検定によるものである。

5 考察

専門用語は必ずしも体系的に命名されているわけではない[2]。構成要素間の係り受けが複数考えられることもある。専門家であっても内部構造の定義は簡単ではない。「慢性肉芽腫性疾患」で「慢性」の係り先は「肉芽腫」「疾患」のどちらも正解と受け入れられるものであろう。

また、タグ付け作業、即ち作業者が考えるその言葉の意味を係り受け関係へ変換する際には構文木や係り受け関係の考え方を理解していることが必要である。今回トップダウンに語を分割する手続きだったこともあり、言語学に馴染みの無い人には直観的でない場合がしばしば見られた。例えば「甲状腺機能亢進」を分割する場合、「甲状腺（の）機能（が）亢進（する)」なので、最初に「甲状腺機能+亢進」と分割すべきだが、「甲状腺（の）機能亢進」と捉えてそこで分割してしまいがちであった。トップダウンな方法の利点は、構成要素として出現した語が既に内部構造を定義されていた場合に同じ定義を繰り返さなくて良いという点である。この利点を生かしたままボトムアップ的に作業できる環境が望ましいと言えよう。

実験の結果はベースラインを大幅に上回っているものの、実用には耐えない精度であった。今回は共起情報など語（構成要素）についての素性を利用していない。LSD 掲載語の一部には意味カテゴリが付与されているので、これを取り入れられるよう工夫したい。

6 おわりに

専門用語の内部構造の表現方法としてラベル付きの構文木、文字単位の係り受けを提案した。また SVM を用いて文字単位の係り受け解析を試みた。対象としたのは疾患名を始めとした生命科学分野の語である。文字対に対する係り受け関係の有無の判定精度は 93.7%であり、この識別器を用いた内部構造解析の精度は 55.6%であり、充分な精度とは言い難い。今後は他の手法を視野に入れながら、識別器の精度向上と内部構造解析のアルゴリズムの改良を進める。

謝辞

本研究の一部は，文科省統合データベースプロジェクト「ライフサイエンス分野の統合データベース整備事業」の支援を得て行われました．また，ライフサイエンス辞書を提供していただいた京都大学金子周司教授に感謝します．

参考文献

[1] 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).

[2] 大島智夫．日本の医学用語についてのおぼえ書き．専門用語研究．No.1, pp.18-21 (1990).

[3] 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).

[4] 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).

[5] 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)