日本語医学専門用語の構造解析

○小　山　照　夫
Teruo KOYAMA
大　江　和　彦
Kazuhiko OHE

Stractural Analysis of Japanese Medical Terms

Abstract

Describing medical data or knowledge, natural language representation plays quite important roles. In machine processing of information described in natural language, concept dictionaries that describe meaning and relationship of various terms are very important. Because most of medical terms, like terms in other specialized fields, can be regarded as composite terms, compiling a medical concept dictionary, it seems effective to prepare elementary terms and composing rules . In this paper, the authors, from the viewpoint of elementary terms and composing rules, discuss about inferring conceptual category of medical terms in a large scale dictionary. Discussions in this paper will concentrate to terms that represents various site, organ or tissues in human bodies.

１．はじめに

医学・医療の分野では、医療行為の記録としての、疾患や病因、あるいは患者に関して観測された症状や検査結果、治療手段としての投薬・手技などのさまざまな情報が、主として自然言語表現を用いて記録されている。また、実際にさまざまな医療行為を遂行していく上で必要とされる知識も、しばしば自然言語表現によって記述されている。

自然言語によって記述された情報の機械処理を行うにあたって、まず第一に必要とされるのは、概念記述の基本となる用語、特に、それぞれの対象分野における専門用語について、その表す概念の意味カテゴリーを明かにするとともに、概念間の相互関係を整理した用語辞書を整備することである。このような概念辞書整備の試みとしては古くから、シソーラスに関する研究が行われている。さらに近年になって、知識ベースの利用を含む医療情報システムの高度活用とも関連して、ＮＬＭにおけるＵＭＬＳ[1] など、新しい概念辞書構築の試みも行われている。概念辞書の効率的な構築のためには、編集作業を効率化すると同時に、概念および概念間の相互関係を判断する材料を提供できる枠組みを用意することが望まれる。このような支援情報を提供する枠組みとして、合成用語としての専門用語の構造に注目することが考えられる。

医学・医療をはじめとするさまざまな専門分野において用いられる専門用語の多くは、より基本的な概念を表す語を合成することによって得られる合成語の形を取る。このような合成語として構成された専門用語は、多くの場合、合成にあたって用いられた元の要素語の意味を反映しているということができる。合成語としての専門用語の意味する概念を、その合成にあたって用いられた、より基本的な用語の属する概念カテゴリーの集合として捉えることは、先に述べた概念辞書編集にあたっての有力な支援手段を提供する可能性がある。また、このような表現を概念の意味を記述する手段として直接利用することにより、機械による専門用語間の意味的マッチングを効率的に行わせることも期待できる。 以下では、この問題に関して筆者らが行った検討の結果のいくつかについて述べる。

２．用語辞書と部位関連語

今回の検討では基本となる用語集として、約２５万語の英語医学専門用語集を日本語に翻

訳したもの（ＭＥＩＤ：医学用ＡＩ電子化辞書、医学用ＡＩ電子化辞書研究会編、日商岩井、１９８９）を用いた。これらの用語の内、日本語訳が純粋に漢字のみからなるもの約１８万語から、英語表現の同意語等に由来する、文字列としてみた場合の重複を除いた、約１３万語を対象とした。今回用いた辞書は比較的大規模な物であるため、対象とする用語の範囲を、人体各部、各種臓器、あるいは局所的な組織系を表す語（人体部位関連語）に限定して検討を行った。

解析を進める上でまず第一に問題となるのは、どのような語を要素語として考え、その語にどのような概念カテゴリーを割り振るかである。これは、現時点ではある程度暫定的な要素語を設定して、これに基づく検討を行わざるを得ない。これまでに行ったいくつかの検討から、人体部位関連用語として基本的なものがいくつか明らかになっている [2]。これらの用語を中心に、約2,200の人体部位関連語に対して、これらを合成する要素語となりうるもの約1,000を暫定的に決定し、これに基づいて検討を行った。

これまでの検討の中で筆者等は、人体部位関連語については、要素語からの合成規則として、［部位＋部位－＞部位］というものが想定できることを指摘してきたが [2]、従来の検討では同一レベルで部位として取り扱ってきた要素語について、今回は、より詳細な分類を試みた。具体的には、次の二つのカテゴリーを想定している。

a.臓器としてまとまりのあるもの、あるいは人体内での位置が明確に同定できるもの（脳、胃、肝臓、肺、頭など）
b.一応器官としてまとまっているが、その具体的位置を同定しにくいもの（血管、神経、骨、筋など）、または、器官というよりはむしろ局所的な組織系を示すもの（粘膜、皮質、腺など）

これらに加えてさらに、人体部位に関連する合成語の構成に関連する要素語として、次のカテゴリーのものを考える。

c.直接的あるいは比喩的に形状を表すもの（円柱、楕円、鞍、蝶番など）
d.空間的相対位置関係を示すもの（上、下、間、囲、端など）
e.機能に関するもの（運動、受容、呼吸など）
f.その他

解析の準備として、まず、選び出された複合語を、暫定的な要素語に分解した。分解にあたっては、最長一致によるパターンマッチを行っている。このようにして得られた分解結果について、以下の検討を行った。

２．１．カテゴリー別出現頻度

合成語の分解結果について、まず、あらかじめ選ばれた人体関連語について、どのようなカテゴリーに属する要素語が高い頻度で出現するかを調査した。要素語は常に単一の概念カテゴリーに属するとは限らず、ある程度の多義性がある場合もある。ある要素語が一つの合成語の中でどのような概念カテゴリーとして取り扱われているかは、前後の文脈から決定される。現時点では、多義性を持つ要素語が、特定の場合にどのカテゴリーとしての役割を担っているかを自動的に判断するのは困難であるため、近似的に、多義性のある語の所属しうるカテゴリーがｎ個存在する場合には、それぞれのカテゴリーに対して１／ｎの出現があったとして集計を行った。その結果、カテゴリーの出現頻度の高いものとして、順に、b.:2,792、a.:759、c.:334、e.:188、d.:131 という結果を得た。

２．２．連続する二つの要素語のカテゴリーの間の関係

合成語を要素語に分解した結果について、カテゴリー別の要素語の出現頻度に次いで興味のある問題として、連続する二つの要素語のカテゴリー間の組み合わせに、何らかの傾向があるかどうかである。これは、どのようなカテゴリー接続パターンが出現するかを集計することによって明かにすることができる。前節と同様な、多義性の補償を行った結果、かなり高頻度に出現しているパターンとして、次のものが得られた。

1.b→b　汎在器官ないし局所組織系同志の接続
2.a→b　部位・臓器から汎在器官ないし局所組織系への接続
3.e→b　機能要素から汎在器官ないし局所組織系への接続

また、頻度の面で、これらに次ぐものとして、次のものも有力である。

4.c→b　形状要素から汎在器官ないし局所組織系への接続
5.d→b　空間的相対位置関係から局所組織系への接続

今回は、最初の３つのパターンについて、検討を行った。

２．１．で用いた要素語辞書を用いて、元の用語集に含まれるすべての用語を、可能なかぎり分解することを試みる。もちろん、要素語辞書は、現段階では完全なものではないため、用語の完全な分解は不可能であることが多いが、それでも、現在要素語辞書に登録してある文字列を副文字列として持つ用語については、可能な範囲で妥当な切り出しを行うことができる。ただし、もちろん、このような分解が常に正しいという保証はない。

このような分解を行った上で、連続して要素語辞書内の文字列が出現する場合について、上記の３つのパターンが出現するもののみを抽出する。このような用語には、今回目的としている部位関連語を表す文字列を含むものも、もちろん相当な割合で含まれているが、いくつかの理由から、部位関連語として不適当なものも含まれている。このような不適当な語が含まれる代表的な理由として次のものが挙げられる。

１．文字列として扱う場合の、分解の多義性（他の要素語の一部を誤って取り込む）
２．要素語自体に解釈の多義性が存在する（特に機能を表す要素語の係り方）
３．合成規則が必ずしも正しくない

これらの問題は、将来、要素語辞書をより整備し、合成規則を洗練することにより、改善できると思われる。しかし、さしあたっては、これらの中から、人体部位関連語として妥当性の高いものを選択できる方法かどうかにも興味がある。

合成語としての部位関連語は、必ずしも先に挙げた３つのパターンに従って二つの要素に分解できるものとは限らないが、元の用語集に二つの要素を合成したものに相当する文字列が含まれるならば、それは、かなり高い可能性で、部位関連語となっていることが期待できる。また、先にも述べたようにこれらがさらに、形態異常や手技を表す要素語に接続する場合、部位関連語である可能性はさらに高まると考えることができる。

実際に最初の約１３万語を、可能なかぎり要素語に分解するという方法によって、この中から、指定されたパターンが含まれる可能性があるものを選択した。これらの中から、先に述べた約 2,000語の、既知の部位関連語に関係するものを除外すると、約10,000語が目的とするパターンを含むことがわかった。これらの中には、目的パターンの前後に別の文字列

をともなっているものも存在するため、目的パターンの文字列だけに限って見れば、約、4,000の文字列が得られることとなる。

これらのうちで、文字列そのものが、元の用語辞書に存在し、なおかつその後ろに「腫」、「癌」などの形態異常要素語または、「摘出」、「切除」などの、手技に関する要素語の来るものを選び出した。この結果、116の文字列が得られたが、そのうち、104については、人体部位関連語とみなすことができた。残りのものとしては、「開口」などの、独立して部位関連語として扱うことに疑問のあるもの、「結核」（「結」を機能とみなし、「核」を局所組織とみなした結果）のように、分解の誤りと要素語の多義性に関わるものや、先に不適切なものの3番目に述べた並置関係にあるものなどがあった。

並置関係として現れるものをもう少し詳細にみると、その多くは［組織系－＞組織系］という構造を持っている。一方、合成語として［組織系－＞組織系］という構造が現れる場合のいくつかを調べてみると、多くの場合、先に来る組織系が後に来るものを空間的に包含している形を取っている。これは、ある意味では自然な関係であり、特定の組織系が、空間的にどの部分に位置しているかを示しているとも考えられる。将来的にはこのような関係を用いることにより、並置関係については、かなりの程度除外できる可能性があると考えている。

以上の結果は、既に約 2,000語を、人手により人体部位関連語として選び出したものにさらに付け加える形でのものであり、要素語への分解と、その結果に含まれる典型的パターンが、部位関連語の推定に相当程度有効性を持つことを示すものといえるであろう。

5．まとめと今後の課題

今回の検討では、日本語医学専門用語に関連して、人体部位・臓器、局所組織など、人体部位関連語を中心に用語の持つ構造に関する検討を行った。

具体的には、暫定的に選択した人体部位関連語に対して要素語となりうる語に関する辞書を用いて、あらかじめ選択された人体関連語について、カテゴリーごとの各要素語の出現頻度と、接続の傾向を検討した。またこの結果に基づき、暫定的な要素語を用いて用語集に含まれる文字列を分解し、その中から、要素語の並びに特定のパターンを持ち、さらに特定のカテゴリーの要素語に接続するものについて検討した結果、このような特定のパターンが、人体部位関連語の推定に有効であることを示した。

今回の検討の結果として、人体部位関連語の推定について、ある程度有効な方法を明かにすることができたと考えている。また、人体部位関連の用語について、その細分類の必要性を明らかとするとともに、これに関連するいくつかの要素語のカテゴリーについて、その役割を明らかにすることができたと考える。

今回用いた要素語辞書や合成規則は試験的な意味合いが強く、まだまだ不完全なものである。今後は要素語辞書や合成規則をさらに充実・洗練するとともに、人体部位以外の概念カテゴリーに関しても、検討の範囲を広げて行くことを予定している。具体的には、人体各部の機能や、疾患概念を表す用語についても検討を進めて行く予定である。

参考文献

1.Lindberg, D. A. B., et. al.: Current Status of the UMLS Project, Proc. SCAMC'90, p.121-154 (include 7 papers), 1990

2.小山照夫、大江和彦：　日本語医学専門用語の構造解析、第13回医療情報学連合大会論文集、p.129-132、Nov. 1993、東京

学術情報センター研究開発部　Dept. R&D National Center for Science Information Systems
東大病院中央医療情報部　Hospital Computer Center, University of Tokyo Hospital