実践医療用語の語構成要素抽出の試み

著者	内山 清子, 岡 照晃, 東条 佳奈, 小野 正子, 山崎 誠, 相良 かおる
雑誌名	言語資源活用ワークショップ発表論文集
巻	3
ページ	463-467
発行年	2018
URL	http://doi.org/10.15084/00001680

実践医療用語の語構成要素抽出の試み

内山 清子（湘南工科大学工学部コンピュータ応用学科）†
岡照晃（国立国語研究所コーパス開発センター）
東条佳奈（目白大学社会学部社会情報学科）
小野正子（西南女学院大学保健福祉学部）
山崎誠（国立国語研究所言語資源研究系）
相良かおる（西南女学院大学保健福祉学部）

Extracting of Word Constituents contained in Medical Terms

Kiyoko Uchiyama (Dept. of Applied Computer Sciences, Sonan Institute of Technology)
Teruaki Oka (Dept. Corpus Studies, NINJAL)
Kana Tojo(Faculty of Studies on Contemporary Society, Mejiro University)
Masako Ono(Faculty of Health and Welfare, Seinan jo Gakuin University)
Makoto Yamazaki (Dept. Corpus Studies, NINJAL)
Sagara Kaoru (Faculty of Health and Welfare, Seinan jo Gakuin University)

要旨

医療現場で用いられる電子カルテなどの記録文書（医療記録）に専門用語としての医療用語が大量に含まれている。医療記録に記載された言語情報を正確に理解・活用するためにはこれらの医療用語の理解が必要となる。

医療記録に含まれる語には、複数の語からなる複合語や臨時一語も多く、これらは、病名、身体の部位名、処置名、薬剤名等、様々な用語から構成されている。しかし、現在はこの語構成要素の組み合わせのパターンや語構成要素間の関係などが曖昧である。

そこで、本研究では複数の語からなる実践医療用語の語構成要素の抽出を試みた。語構成要素の条件を独自で定義した後、ComJisyoV5、と今後公開予定の V6 の登録候補語に対象として、MecabMeCab 0.996 と UniDic-cwj-2.2.0 を利用して形態素解析を行った。分割された単語の品詞情報を手がかりにして、単一単位となり得る品詞列を抽出した。次に抽出した候補リスト以外に語構成要素となる品詞列があるかについて検討を行った。

１．はじめに

医療記録には、専門用語に加え、略語や隠語など独特な表現が含まれる。この記録データを電子カルテシステムの普及により、施設内での共有や、大量の医療記録データを二次利用する研究なども増加してきている。しかし、実際に医療記録に記載された言語情報を正確に理解・活用するためには、専門用語を含む複合語や臨時一語などの用語を適切に抽出することが不可欠である。本研究では、複数の語からなる実践医療用語に特化した形態素解析辞書の構築を目的とし、実践医療用語を構成している語構成要素を抽出することを試みた。

† uchiyama@sc.shonan-it.ac.jp

日本語学において、複合語の構造を明らかにする「語構成論」について、斎藤(1996)は語構成要素から語が成立するまでの一連のプロセスの内容を明らかにする研究分野として定義している。斎藤(2004)では、従来の語構成論では複合語が語構成論の対象となっているのに対して、単純語も複合語も同じく語構成論の対象となるという立場に立っている。また、単語化、語構成要素、語構成要素間の関係の 3 つの観点から語構成論を論じている。

本研究における語構成要素とは、専門用語を理解する上で分割できる最小単位の語に相当するものであると考えている。つまりすでに辞書に登録されている 2 文字や 3 文字の名詞や動詞などの単語だけでなく、複数の単語や形態素が結合した単位も含んでいる。この語構成要素を確定することで、語構成要素の意味が理解できていれば複数の語構成要素からなる臨時一語や複合語などの専門用語を理解するための支援になる。今回は従来の形態素解析で処理した結果をどの程度活用できるかを検討するために分析を行う。以下、2 章において分析対象としたデータについての説明を行い。3 章に抽出方法、4 章に抽出結果を述べ、その後考察と今後の課題について記述する。

２．対象データ（ComJisyoV5）

医療従事者用の臨床記録文書（看護記録、プログレスノート、医療経過記録など）を解析するための支援として 2008 年から形態素解析辞書 ComeJisyo を作成し、2013 年 11 月に ComeJisyoV5（登録語数 77,760 語）を公開している。本研究では、文字長が 2 文字以上の用語を対象とし、非公開の研究用見出し語データ 52,974 語と ComeJisyo V5-1 および公開予定の ComeJisyoV6 の登録語の併せて 109,721 語で重複している 31,162 語を抽出し、本研究における語構成要素候補とした。この候補は 2 文字から 22 文字から成る医療用語が含まれているため、語構成要素と複数の語構成要素を組み合わせた複合語になっている。この用語のうち、複合語がどのような語構成要素から成り立っているのかを調べるために、まず語構成要素自体を独自に定義し、抽出することが必要となる。

３．抽出方法

まず、語構成要素の候補となりうる条件を設定した。語構成要素の候補とならないものとして臨時一語を含む複合語の認定条件をあげた上で、その条件に該当しない、かつ機械的に抽出することが可能な条件を考えた。

石井(2007)によると、「臨時一語の認定条件」は、1)複数の単語が臨時的に結びついたもの、2)複合語、3)もとの単語列に復元することができるものとしている。反対に、臨時一語と判定されないものは、①固有名、②組織名・役職名、③ときの表現、④地名、⑤数量に関する表現としている[1]。このように臨時一語の認定条件に該当しないものが語構成要素の候補になると定義し、形態素解析結果から品詞を指定して抜き出す条件を考えた。

語構成要素候補リスト 31,162 語に含まれる用語を対象として MecabMeCab 0.996 と UniDic-cwj-2.2.0 で形態素解析を行い、品詞を付与した。

表 1 に形態素結果語の語構成要素リストの構成要素数を示す。

表 1 構成要素数

構成要素数	該当単語数
1	4062
2	8927
3	7789
4	5637
5	2878
6	1178
7	453
8	168
9	51
10	14
11	4
12	1

表 1 の通り、構成要素が 1 つからなる用語は 4062 語含まれており、この品詞は記号が 1 語と形状詞が 12 語、後は全て名詞となっていた。このことから形状詞や名詞は単独で語構成要素の条件として設定することが妥当であると判断した。そこで、「臨時一語の認定条件」や臨時一語と判定されないものを参考にしながら、語構成要素候補として以下の品詞列の条件を設定した。

①品詞列の条件
単一語 :「名詞」または「形状詞」のもの
二語以上 :「名詞」＋接尾辞」、「形状詞」＋「接尾辞」の並びになっているもの、「記号のみの組み合わせ」
「名詞」＋「名詞」の 2 文字
②①以外で文字数 3 文字以下のもの

このように①と②に該当する用語を抽出し、語構成要素として認定可能かどうかを分析していく。

４．抽出結果

3 章の抽出ルールに従って語構成要素候補を抽出した結果を表 2 に示す。

表 2　抽出ルール別該当数

	抽出条件	13,849
①	名詞	4062
	形状詞	12
	名詞｜接尾辞	1418
	形状詞｜接尾辞	16
	名詞｜名詞　の 2 文字	74
	記号のみ	188
②	3 文字以下	1299
	条件①と②の合計	7327

表 2 に示した通り、構成要素の品詞として名詞がもっとも多く含まれていた。また、「名詞＋接尾辞」の組み合わせも多く出現していたことや、記号の連続も専門用語の一部を構成しているものがあるなど、条件として①は適切であったと考えられる。

次に①の条件を満たさない 3 文字以下の単語であるが、①の条件を除いた後に分析すると名詞、形状詞、記号を含む用語が全て対象外となってしまう。残った単語を見てみると感動詞、助詞、助動詞、動詞、副詞など解析誤りを含むものが多かった。そこで、再度 2 文字以上 3 文字以下の用語には元々どのような品詞列が含まれているかを調べた。

表３　文字数 3 以下の用語の品詞列

品詞列	用語数	品詞列	用語数
名詞/名詞	1012	記号/接尾辞	16
名詞/名詞/接尾辞	55	名詞/記号	11
名詞/接尾辞/接尾辞	34	記号/名詞	10
接頭辞/名詞/接尾辞	30	名詞/名詞/名詞	10
名詞/接尾辞/名詞	29	形状詞/名詞	8

表 2 や 3 で分かる通り、「名詞＋接尾辞」の品詞列がもっとも多いことから抽出ルールにこの条件を入れた妥当性が確認できる。文字数 3 文字以下という条件をつけると 2 文字で「名詞＋名詞」や「接頭辞＋名詞」の用語は一つの語構成要素とすべきものを多く含んでいる。3 文字で「一側性」などの「名詞＋名詞＋接尾辞」、「一次的」「一次性」などの「名詞＋接尾辞＋接尾辞」、「両価性」などの「接頭辞＋名詞＋接尾辞」といった品詞列はこれも一つの語構成要素と認定しても良いものである。このように 3 文字以下の用語から長単位の品詞列をまず確定させてから、残りの用語について分析をするのが適切な手順であったのではないかと考察する。

同様の問題として、名詞の中には一文字単語である「群」「型」などが 888 語も含まれていたが、これらを単独で語構成要素とすることは難しい。

５．考察

ここまで、医療用語の語構成要素を選定するにあたって、まず品詞列をてがかりとした条件を設定して抽出を試みた。その結果として「名詞＋接尾辞」「形状詞＋接尾辞」が語構成要素候補としてもっとも適した条件であった。一方で、文字数と品詞列の対応をさせることによって、条件を細かく設定すればより効率的に抽出できるのではないかと考えられる。その中で、抽出条件の優先順位も考慮する必要がある。たとえば、まず長単位の品詞列を適応させ、その語構成要素を一単位と設定し、次に大雑把な品詞列の条件に当てはまるものを抽出していく手順を考えていくべきである。

また、分析している中で、出現頻度や連接確率などを考慮にすればより効率的な抽出が可能になるのではないかという結論に達した。各語構成要素が用語全体のどの位置で出現しやすいかという分析も試みたが、語構成要素が確定していない中で分析することが難しく、統計的にまとめることができなかった。しかし、確実に用語の頭や終わりに出現する語構成要素や、複数の語構成要素が組み合わさって長い用語を構成する場合には、その順番もある程度規則性があることなどが見て取れた。

６．まとめと今後の課題

本研究は、医療現場で作成される医療記録に記載された言語情報を正確に理解・活用するために必要となる、専門用語を含む複合語や臨時一語などの用語を構成している語構成

要素を適切に抽出することを試みた。形態素解析結果の品詞列を手がかりとして、「名詞＋接尾辞」「形状詞＋接尾辞」「名詞」単独、「形状詞」単独、複数の「記号」列からなる用語を分析し、語構成要素と認定するに適切な条件であることを確認した。

今回は、抽出条件を中心に検討していたため、手順や、条件の優先順位を整理することができなかった。今後は抽出条件と手順を明確にし、更に語構成要素間の連接頻度や結合関係などを分析していきたい。

謝　辞

本研究は JSPS 科研費(18H03499)の助成を受けたものです。

文　献

石井正彦(2007)，現代日本語の複合語形成論，ひつじ研究叢書，ひつじ書房.

斎藤倫明(1996)，現代日本語の語構成論的研究−語における形と意味−，日本語研究叢書，ひつじ書房.

斎藤倫明(2004)，語彙論的語構成論，ひつじ研究叢書，ひつじ書房.