ポスター発表

[PB] ポスター B

2020年6月6日(土) 09:00 〜 16:30 ポスター会場(2) (e-poster)

[PB-30] 実践医療用語の語構成解析
Word component decomposition and semantic composition analysis of medical compound words

*相良 かおる[1]、小野 正子[1]、東条 佳奈[2]、麻 子軒[2]、山崎 誠[3] （1. 西南女学院大学、2. 大阪大学大学院文学研究科、3. 国立国語研究所）

*Koru Sagara[1], Masako Ono[1], Kana Tojo[2], Ma Tzu-Hsuan[2], Makoto Yamazaki[3] （1. Seinan Jo Gakuin University, 2. Osaka University, 3. National Institute for Japanese Language and Linguistics）

実践医療用語の語構成解析

相良かおる[*1], 小野正子[*1], 東条佳奈[*2], 麻子軒[*2], 山崎誠[*3]
[*1] 西南女学院大学, [*2] 大阪大学, [*3] 国立国語研究所

Word component decomposition and semantic composition analysis of medical compound words

Kaoru Sagara[1)] Masako Ono[1)] Kana Tojo[2)] Ma Tzu-Hsuan[2)] Makoto Yamazaki[3)]
1)Seinan Jo Gakuin University 2) Osaka University
3)National Institute for Japanese Language and Linguistics

抄録: 本報告では、分かち書き用実践医療用語辞書 ComeJisyoSjis-1 の登録語から選定した合成語 2,655 語より語構成要素を抽出する方法と得られた語構成要素について述べる。具体的には、国立国語研究所の短単位自動解析用辞書 UniDic により、合成語を斉一な言語単位(短単位)に分割した後、付与された品詞情報を基に機械的に連結し、分割する。これらを意味的な単位にまとめあげ、語構成要素列を求めた後、医療的な観点から意味的かつ統語的に妥当な語構成要素列を求めるという方法をとった。

本手法により、意味的および医療的語構成要素列から 2,620 種類の語構成要素が得られた。

これらの語構成要素は、合成語を含む医療記録文の情報検索に、また、利用目的に適した単位での分かち書き用辞書の作成への利活用が可能である。

キーワード 実践医療用語、合成語、語構成解析、語構成

1. はじめに

医療記録文には、専門的な意味を持つ合成語が、多く含まれている。

語の結合に注目した、合成語を構成する語(語構成要素)の結合パターンが明らかになれば、用途に適した語単位での分かち書き用辞書の作成に、また、機械学習用の学習データとして、利活用できる。

本発表では、一般的な語を含む専門的な意味を持つ合成語より、語構成要素を抽出する方法とその結果得られた語構成要素について述べる。

2. 用語の定義

語構成要素

合成語を構成する要素で、本報告における語構成要素を、「医療の観点から意味的に分割可能な語をすべて語構成要素とする」と定義する。

合 成 語: 先天性脳性麻痺
語構成要素: 先天性、脳性麻痺、脳性、麻痺

語構成要素列

合成語を構成する語構成要素の順序組

合 成 語: 先天性脳性麻痺
語構成要素列:先天性／脳性麻痺

短単位

短単位とは、国立国語研究所が言語の形態的側面に着目して規定した斉一な言語単位である。現代語において意味を持つ最小単位を規定した上で、最小単位を短単位の認定規定に基づき結合させる、または結合させないことにより、認定される。

合 成 語: 先天性脳性麻痺
短 単 位: 先天、性、脳性、麻痺

3. データ

医療の知識を持たない共同研究者による意味的な分割を考慮し、分かち書き用実践医療用語辞書 ComeJisyoSjis-1[1]の登録語 111,664 語より一般的な語(『分類語彙表 増補改訂版』[2]収録の語)を含む合成語 7,139 語から筆者などが任意に選んだ 2,655 語を本解析データとする。

4. 方法

step.1 機械的分割(機械的順序組)

A) 形態素解析器 MeCab0.996[3]と見出し語約 87 万語の解析用辞書 UniDic[4]により、合成語を短単位に分割
B) 付与された品詞ラベルが「形状詞」または「記号」の場合、品詞ラベルを「名詞」に変更
C) 「接尾辞」の語を直前の語に連結
D) 「接頭辞」の語を直後の語に連結
E) 連続する「カタカナ」のみの語を連結

step.2 意味的分割(意味的要素列)

共同研究者 5 名により意味的に妥当な語構成要素の順序組(以下、語構成要素列)を求める。各専門領域は、日本語学が 3 名、情報科学が 1 名、看護学が 1 名である。

step.3 医療的観点による分割(医療的要素列)

臨床看護の経験者 1 名により医療の観点からみて意味的にも統語的にも妥当な語構成要素列を求める。参考にした辞書などを以下に示す。

- 医学書院 医学大辞典
- 医学英和辞典 第 12 版
- 医学書院 看護大事典 第 2 版
- 南山堂 医学大辞典 第 20 版
- ステッドマン医学大辞典 改訂第 6 版
- ブリタニカ国際大百科事典 2019
- 広辞苑 第 7 版

5. 結果

1) 語構成要素列

Table 1 語構成要素列の概要

	機械的	意味的	医療的
平均	2.9	2.4	2.2
中央値	3	2	2
最小値	1	1	1
最大値	8	6	6

合成語 1 語あたりの機械的要素列における要素数の中央値は3語、最大値は8語、意味的要素列と医療的要素列における要素数の中央値は 2 語、最大値は 6 語であった。

Table 2 合成語 1 語あたりの語構成要素数

要素数	機械的		意味的		医療的	
1	100	4%	225	8%	210	8%
2	932	35%	1,317	**50%**	1,686	**64%**
3	1,017	**38%**	878	33%	673	25%
4	445	17%	198	7%	75	3%
5	108	4%	28	1%	10	0%
6	32	1%	9	0%	1	0%
7	9	0%	0	0%	0	0%
8	2	0%	0	0%	0	0%
無し[注)]	10	0%		0%		0%
計	2,655	100%	2,655	100%	2,655	100%

注）Step1A)～E)で語構成要素が求まらなかった語

機械的に語の順序組が得られなかった合成語が 10 語あった。その内訳は、UniDic 辞書の未知語に依るものが 6 語（「迷もう」「細網」「顔位」「額位」「乏尿」「醜形」）、品詞の誤りに依るものが 3 語（「ばち[副詞]／指・爪」、「第一[副詞]」）、誤解析によるものが1語（「が[助詞]ま[感動詞]／腫」）であった。

Table 3 語構成要素列の比較

	一致		不一致		計
機械的:意味的	1,584	59.7%	1,071	40.3%	2,655
機械的:医療的	1,361	51.3%	1,294	48.7%	2,655
意味的:医療的	2,023	76.2%	632	23.8%	2,655

品詞を基にした機械的要素列と、意味を基にした意味的要素列との一致度は 59.7%、医療的要素列との一致度は 51.3%であった。一方、意味的要素列と医療的要素列の一致度は 76.2%であり、23.8%が不一致であった。

機械的、意味的、医療的要素列が全て一致した要素列に「ノロウイルス性／腸炎」、「壊疽性／丘疹状／結核疹」などがある。意味的要素列と医療的要素列が異なる例を以下に示す。

意　味　的: おとがい／神経／麻痺
医　療　的: おとがい神経／麻痺／

意　味　的: 脛骨／動脈／損傷
医　療　的: 脛骨動脈／麻痺／

要素列 3 種共に異なる例を以下に示す。

機　械　的: 胸／管内／頚／静脈／吻合／術
意　味　的: 胸管／内頚静脈／吻合術
医　療　的: 胸管内頚静脈／吻合術／

2) 語構成要素数

Table 4 語構成要素数の概要

	機械的[注)]		意味的		医療的	
	異なり	延べ	異なり	延べ	異なり	延べ
要素数	1,996	7,595	2,143	6,472	2,271	5,957

注） 細分割できない合成語 10 語の要素は含まれない

本手法により、意味的要素列に含まれる 2,143 語（異なり）と医療的要素列に含まれる 2,271 語（異なり）から重複を除いた 2,620 語の語構成要素が得られた。

6. 考察

本手法において、UniDic 辞書による機械的分割の手法は、一般的な語を含む合成語を対象にした場合、誤解析が少なく（3 語）有用であった。なお、紙面の都合で割愛したが、本合成語の 95%（2,358 語）が医療情報システム開発センター(MEDIS-DC)の病名マスターの索引語と一致する専門用語である。

一方、意味的要素列への分割は、分割規則を定め共同研究者で分担して分割した後、担当データを変えて再度見直したが、若干の誤分割が含まれている。また、医療的要素列の分割者は臨床看護経験者であり、共同研究者に医師は含まれていない。従ってそれぞれの要素列について、見直が必要と考えている。

医療的要素列については、一意に定まらない要素列「乳汁分泌／抑制」と「乳汁／分泌抑制」が見つかった。電子版『南山堂医学大辞典 第 20 版』の部分一致検索結果、「成長ホルモン／分泌抑制」、「胃酸／分泌抑制」、「酸／分泌抑制」より、頻度による結合の強さから医療的要素列を求めると「乳汁／分泌抑制」となる。しかし医療的要素列の分割者および臨床看護経験を持つ筆者からの「『乳汁分泌／抑制』の方が、違和感がない」との意見に、実際の利用者である助産師に確認したところ、「乳汁分泌／抑制」との回答を得た。このことは、共起頻度などで機械的に求めた医療的要素列が、実際の利用者にとってかならずしも妥当ではないことを示唆している。

7. 結語

医療記録に含まれる合成語の語構成の解明と語構成要素の抽出を目的に、①UniDic 辞書を用い、機械的に分割を行い、②意味的要素列を求めた後、③医療的要素列を求めた。その結果、一般語を含む合成語 2,655 語より、2,620 語の語構成要素が得られた。

自然言語処理において、分かち書きは最も重要な処理である。本分析で得られた結果を基に、意味的にも統語的にも妥当な単位で分かち書きできる辞書を作成・公開する予定である。

謝 辞

本研究は JSPS 科研費 JP18H03499 の助成を受けています。

参 考 文 献

[1] ComeJisyo：https://ja.osdn.net/projects/comedic/
[2] 国立国語研究所：分類語彙表 増補改訂版，大日本図書，2004.
[3] MeCab： https://taku910.github.io/mecab/ (参照 2020-03-22)
[4] UniDic：https://unidic.ninjal.ac.jp/ (参照 2020-03-22)