ComeJisyoUtf8-3 の誤解析調査

ー看護師・助産師・管理栄養士国家試験問題文の語分割ー

相良かおる
西南女学院大学　保健福祉学部
sagara@seinan-jo.ac.jp

概要

実践医療用語辞書 ComeJisyoUtf8-3 の語分割の精度を，看護師・助産師・管理栄養士国家試験問題文を解析し調べた．

その結果，全見出し語が 48,191 語，その内，未知語成分が 1,108 語（2.3%），これらから成る未知語が 382 語（異なり 320 語），そして誤解析（過分割）された登録語の見出し語が 421 語（0.9%），当該登録語が 155 語（異なり 47 語）であった．これらには，数字や記号を含む登録語の過分割が多くあった．

1 はじめに

筆者は，電子化された医療記録データの自然言語処理を支援するために実践医療用語を見出し語とした分かち書き用辞書 ComeJisyoV1 を 2008 年 11 月に公開した．その後，随時更新し，2021 年 3 月公開の ComeJisyoUtf8-3 の登録語数は約 12 万語となった．

なお，門外不出の医療記録に含まれる実践医療用語の語構成や語種構成の実態は不明であるため，語の単位認定の規則を定めず，医療従事者が一つのまとまった語としたものを登録している．

ComeJisyoUtf8-3 の登録語は，看護経過記録，プログレスノート，医師経過記録，看護師・助産師・管理栄養士の国家試験問題文（1998 年度～2019 年度）から抽出しているが，利用状況を調べると，退院サマリ，インシデントレポート等，多様な医療文書の解析に利用されている[1][2]．そして，ComeJisyo のダウンロード数は 2015 年以降増加している（図 1）．

Google Scholar で検索語を“ComeJisyo”として検索すると[i]，通年での検索件数は 41 件，期間を 2015 年～2021 年に絞ると 34 件，“ComeJisyo”と“machine learning”の AND 検索では共に 11 件，“ComeJisyo”と“機械学習”の AND 検索では，10 件と 8 件となり，機械学習に関する研究での利用が増加している．

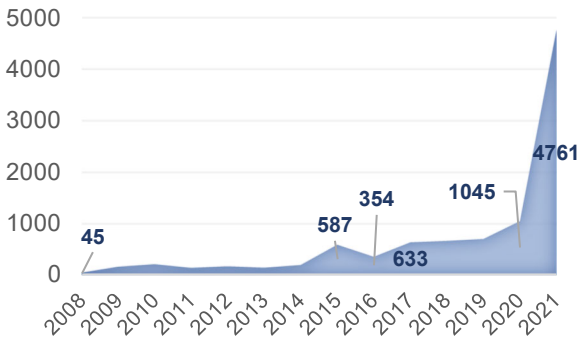

図 1 ComeJisyo ダウンロード数の推移

なお，筆者が ComeJisyo を公開する目的は，医療施設内，または診療科内で蓄積されたスモールデータの自然言語処理の支援である．

一般に単語分割の評価基準には，精度，再現率，F 値が用いられる[3]．

$$\text{精度} = \frac{\text{正しく抽出された単語数}}{\text{システムが出力した単語数}}$$

$$\text{再現率} = \frac{\text{正しく抽出された単語数}}{\text{正解の単語数}}$$

$$\text{F 値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}}$$

しかし，ComeJisyo の登録語には，「びまん性管内増殖性糸球体腎炎ネフローゼ症候群」のような長い語と，その部分文字列である「ネフローゼ症候群」が登録されている．

そして，利用者が必要とする語の単位も，解析する医療文書の種類も分からない．従って，利用者にとって有益な精度や再現率を提示することは困難である．また，実用目的で医療文書を解析する際に

[i] 検索日：2021 年 12 月 27 日

は，精度や再現率の値に関わらず人手により解析結果の確認・修正が行われると推測される．

そこで本発表では，ComeJisyo の具体的な解析誤りを示すことが有益だと考え，看護師・助産師・管理栄養士の 3 種類の国家試験問題の解析結果における未知語と誤解析について述べる．

2 用語の定義

本発表で用いる用語を以下に定義する．

登録語：ComeJisyo または IPA 辞書に登録の語

見出し語：MeCab の実行で出力された 1 語

未全文字列：

以下を総称して「未全文字列」という．

(1) 造語成分：「ー煮」，「蒸しー」等
(2) 接頭語（詞）：「要ー」，「同ー」等
(3) 接尾語：「ー時」，「ー語」等
(4) 異義語と解釈された語：
「よ／だれ」[ii] →「だれ（誰）」等
(5) 古い表現：「安気」等
(6) 不明な文字列：「よ」等

過分割：

以下の 3 種類の分割をいう．

(1) 意味のある語（ComeJisyo または IPA 辞書の登録語を含む）に分割される
(2) 「未全文字列」を含む語に分割される
(3) 数字部分が分割される
(4) 英字文字列の途中で分割される
“overflow” → “over／flow”

誤解析：

本発表では以下を誤解析という．

(1) 登録語が過分割される
(2) 登録語の異表記（半角/全角，大文字/小文字，ヵ/ヶ等）による未知語成分の出力
(3) 句点，読点，記号等の品詞の間違い

3 言語資源の概要

本発表では，2020 年度に実施された第 110 回看護師国家試験[iii]，第 104 回助産師国家試験[ii]，第 35 回管理栄養士国家試験[iv]の問題文（PDF 形式）を UTF-8 形式でデジタル化し，解析データとする．

表 1 に，解析データの文字数，行数を示す．

表 1 解析データの文字数と行数

	看護師	助産師	管理栄養士	計
字数	34,303	18,219	35,087	87,609
行数	1,332	664	1,264	3,260

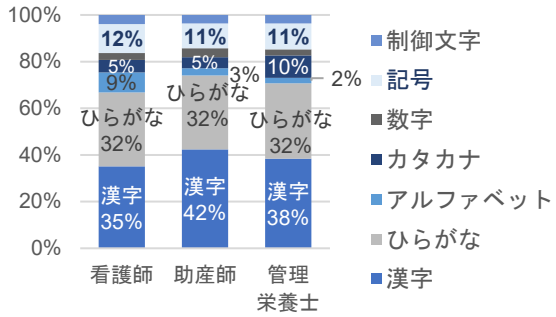

図 2 文字種の概要

図 2 は文字種の割合をまとめたものである．

4 調査方法

4.1 環境

本調査の環境を示す．

形態素解析器：mecab-0.995.exe[v] (MeCab)
システム辞書：IPA 辞書[iv]
ユーザ辞書：ComeJisyoUtf8-3.dic[vi] (ComeJisyo)
ユーザ辞書 2：NEologd.20200910-u.dic[vii] (NEologd)

4.2 方法

調査の手順を以下に示す．

手順 1. ComeJisyo を用いて解析データを解析
手順 2. 未知語を構成する見出し語（未知語成分）を調べる
手順 3. 誤解析された登録語の見出し語（登録語誤解析）を調べる
手順 4. 未知語数と誤解析の登録語数を調べる

5 結果

表 2 は，解析結果の概要である．

「未知語成分」には，ComeJisyo または IPA 辞書

[ii] 「／」は分割位置を示す．
[iii] 厚生労働省 https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp210416-03_04_05.html
[iv] 厚生労働省 https://www.mhlw.go.jp/stf/newpage_17056.html
[v] MeCab：http://taku910.github.io/mecab/#download
[vi] ComeJisyo：https://ja.osdn.net/projects/comedic/
[vii] neologd /mecab-ipadic-neologd：https://github.com/neologd/mecab-ipadic-neologd

の登録語が含まれる（表 3）．例えば，未知語「脳血管障害後遺症」の部分文字列「脳血管障害」は ComeJisyo の登録語である．この場合の未知語成分は，「脳血管障害」と「後遺症」の 2 語となる．

「登録語誤解析」においても，誤解析となった登録語の見出し語が登録語である場合もある（表 4）．例えば「訪問看護ステーション」の見出し語は「訪問看護」と「ステーション」の 2 語であり，これらは ComeJisyo の登録語である．なお，登録語「二人暮らし」の異表記「2 人暮らし」の見出し語は，「2」「人」「暮らし」の 3 語であったが，今回，これらは「未知語成分」ではなく「登録語誤解析」に含めている．

表 2　解析結果（延べ数）の概要

	見出し語 (EOS 含む)	未知語成分 語数	割合	登録語誤解析 語数	割合
看護師	18,480	596	3.2%	201	1.1%
助産師	10,887	277	2.5%	141	1.3%
管理栄養士	18,824	235	1.2%	79	0.4%
計	48,191	1,108	2.3%	421	0.9%

表 3 未知語成分の種別（延べ数）

	ComeJisyo	IPA 辞書	未登録	計
看護師	265 44%	259 43%	72 12%	596
助産師	129 47%	126 45%	22 8%	277
管理栄養士	78 33%	149 63%	8 3%	235
計	472 43%	534 48%	102 9%	1,108

表 4　登録語誤解析の種別（延べ数）

	ComeJisyo	IPA 辞書	未登録	計
看護師	61 30%	71 35%	69 34%	201
助産師	37 26%	38 27%	66 47%	141
管理栄養士	19 24%	27 34%	33 42%	79
計	117 28%	136 32%	168 40%	421

表 5 未知語と誤解析語の語数

	未知語		誤解析	
	延べ	異なり	延べ	異なり
看護師	137	121	79	30
助産師	130	111	50	19
管理栄養士	115	101	26	20
計	382	320	155	47

表 5 は，未知語成分からなる未知語と誤解析となった登録語の語数である．

次章では未知語と誤解析について述べる．

6. 未知語の種類と対処法

医療記録文には，一般的な語も含まれる．今回の未知語 382 語の内，医療用語以外の語は 134 語であり，「よだれ」「いきみ」「おそれ」「ハザードマップ」「過干渉」「給水車」「炊飯器」「大気汚染物質」「育児介護休業法」「共食」「煮魚」「つぶし粥」等があった．医療記録文の語分割には，これらの一般的な語を登録した辞書も必要であろう．

今回，無償で利用でき，新語・固有表現に強く、語彙数が多いとされる NEologd をユーザ辞書に追加して上記 12 語を解析したところ，「大気汚染物質」「育児介護休業法」「共食」「煮魚」「つぶし粥」以外の 7 語は，正しく解析された．

6.1 品詞情報による未知語の生成

連続する名詞の連結

未知語「療養介護」「消化器感染症」の未知語成分「療養」「介護」「消化器」「感染症」は，全て ComeJisyo の登録語であり，品詞は「名詞,一般」になっている．このような場合は，連続する品詞が「名詞,一般」である見出し語を連結することで，未知語の生成が可能である．

接頭語，接尾語の連結

未知語「最優先」の未知語成分「最（接頭詞,名詞接続）」と「優先（名詞,サ変接続）」も IPA 辞書の登録語であり，この場合は，品詞が「接頭詞 名詞接続」の見出し語を後続する品詞「名詞」の見出し語と連結することで未知語の生成が可能となる．また，未知語「手洗い場」の未知語成分「手洗い（名詞,サ変接続）」と「場（名詞,接尾）」も同様に，「名詞,接尾」の見出し語を直前の品詞が「名詞」である見出し語に連結することで未知語の生成が可能となる．

6.2 生成できない未知語

一方，未知語成分が ComeJisyo または IPA 辞書の登録語であっても語の境界誤り等で異義語として解析された場合，品詞のみによる未知語の生成ができない場合がある．例えば，未知語「よだれ」は，「よ（その他,間投）」と「だれ（名詞,代名詞,一般）」に解析され，付与された品詞から未知語を生成することはできない．同様に「お／それ」「かた／まり」「いき／み」も生成が困難である．また，「くん煙した」内の未知語「くん煙」を，未知語成分「くん（名詞,接尾,人名）」と「煙し（形容詞,自立）：ケムシ」から生成することは出来ない．同様に「蒸し器内の」に含まれる未知語「蒸し器」も，未知語成分「蒸し（動詞,自律）」と「器内（名詞,副詞可能）」

からの生成は困難である．

7. 誤解析の種類と対処法

7.1 品詞情報による対処法

過分割された見出し語全てが登録語であるものに「SP／O2」「パンツ／型／おむつ」「訪問／看護業務」「血漿／蛋白質」等があった．これらの品詞は名詞であることから，連続するこれらの見出し語を連結することで，元の登録語を得ることができる．

数字が含まれる登録語は「糖原病／I／型」「3／価／鉄」「2／人」「1／泊」等となった．これらも数字の品詞が「名詞,数」で後続する見出し語が「名詞,接尾」であれば，品詞により連結が可能である．

7.2 その他の対処法

登録語「二人暮らし」の異表記である「2／人／暮らし」や，「血漿タンパク質」「血漿たんぱく質」の異表記である「血漿／蛋白質」が過分割された．これらの表記の揺れは，前処理により表記を統制することで回避することができる．

医療記録文には，記号が多く使われる．本解析データにおいても 11%以上を占めている（図 2）．

ComeJisyo には，「/μL」「mEq/l」「/分」「γ－アミノ酪酸」「n-3 系脂肪酸」など記号を含む登録語が含まれるが，これらの殆どが過分割される．
そして半角記号「/」「-」の品詞は「記号」とはならず，「名詞,サ変接続」となり，品詞規則による訂正は困難である．また，全角「／」「－」は記号となるが，登録語「／分」は，正しく解析される場合と「／」「分」に過分割される場合がある．

同様に句読点の品詞が「記号,句点」または「記号,読点」ではなく「名詞,サ変接続」となる場合がある．従って，品詞情報で合成語を生成する際には，記号や句読点の品詞を確認するか，MeCab の部分解析[viii]を利用する等の工夫が必要である．

8. まとめ

本発表では，3 種類の国家試験問題文を解析データとし，ComeJisyo の解析結果について述べた．

記号，数字を含む登録語を除き，過分割された見出し語の多くは登録語であり，付与された品詞情報から合成語の生成が可能である．なお，解析結果をエディタ等で修正でき，付与された品詞情報を基に合成語を生成するツール GoMusubi_Ver.2.0[ix]を 2021 年 4 月に公開している．

単位認定を定めず登録語の多くが合成語である ComeJisyo の持つ問題に「一貫性のない分割」[3][4]がある．今回は「／分」が，「／」と「分」に過分割されるケースがあった．医療記録文は，主語の省略や体言止めが多く，登録語には，合成語とその構成要素も含まれることから，分割の一貫性は低い．

また，登録語「訪問看護業務」は「訪問／看護業務」と 2 語の登録語に過分割された．しかし意味的な結合の強さからは「訪問看護／業務」と分割するのが妥当であり，これら 2 語も登録語であるが，意味的な結合の強さを考慮した語分割は，ComeJisyo を用いた解析では望めない．

そこで，合成語 7,192 語を意味的に語構成要素に分割し，これらに意味ラベルを付与し『実践医療用語_語構成要素語彙試案表 Ver.1.0』[x]として，2021 年 3 月に言語資源協会より公開している．

今後，これらの合成語に意味的な結合の強さによる語境界を記述した『実践医療用語_語構成要素語彙試案表 Ver.2.0』を公開予定である．

謝辞

本研究は JSPS 科研費 JP21H03777, JP18H03499, JP21300099 の助成を受けたものです．

参考文献

1. 安道健一郎，奥村貴史，小町 守，松本裕治．確度に基づく退院時サマリの分析．情報処理学会研究報告，Vol.2019-NL-240，No.2，pp.1-7，2019.
2. Shinichiroh Yokota，Emiko Shinohara，Kazuhiko Ohe. Can Staff Distinguish Falls: Experimental Hypothesis Verification Using Japanese Incident Reports and Natural Language Processing. Nursing Informatics 2018，Vol.250，pp.159 -163，2018.
3. 工藤拓. 形態素解析の理論と実装. 近代科学社. 2018.
4. 高橋 文彦，颯々野 学．情報検索のための単語分割一貫性の定量的評価．第 22 回言語処理学会年次大会，2016.

[viii] 制約付き解析(部分解析) https://taku910.github.io/mecab/partial.html

[ix] ComeJisyo：https://ja.osdn.net/projects/comedic/

[x] GSK2020-G：https://www.gsk.or.jp/catalog/gsk2020-g/