電子医療記録の分ち書き用ユーザ辞書 ComeJisyo の紹介と単語生起コスト

相良かおる(西南女学院大)、 小野正子(西南女学院大)、小木曽智信（国語研)、 小作浩美（NICT)

1. はじめに

電子カルテシステムを導入している施設では、電子医療記録などのテキスト形式の医療情報（以下、医療情報という）が日々蓄積され、これらには、専門用語、略語、方言など、多種多様な語彙や表現が含まれている。

医療情報をコンピュータで処理し、活用するためには、最初に単語に分割する「分かち書き」作業が必要である。現在、いくつかの形態素解析プログラムが開発・公開されており、これらは専用の辞書を用いて単語分割を行っている。

我々は、医療情報の分ち書きを目的に形態素解析器 Mecab[1]用の辞書 ComeJisyo[2] を作成し、公開している。

本稿では、初めに ComeJisyo の概要について述べ、次に Mecab 用ユーザ辞書に必要な単語生起コスト（以下、コストという）の設定について報告する。

2. 分ち書き（単語分割）の単位

単語間にスペースなどの区切りのない日本語における単語単位の認定は、形態素解析用辞書に登録された単語によって決まる。Mecab 用の辞書には、品詞や読みなどの情報に加えて、登録された単語を含む品詞タグ付きコーパスを基に機械学習により得られたコストが付与されている。

すなわち、この辞書に単語ではなく複合語を登録し、その複合語を含む品詞タグ付きコーパスを用いた機械学習によるコストを付加することで、複合語を単位として分割することが可能となる。

例えば、意味をもつ最小の単位「短単位」を登録した辞書で分かち書きすると、

「バイタル｜サイン｜、｜意識｜レベル｜、｜睡眠｜状況｜、｜歩行｜状態｜観察｜した｜。」

となるが、複合語を辞書に登録することで

「バイタルサイン｜、｜意識レベル｜、｜睡眠状況｜、｜歩行状態｜観察｜した｜。」

のように、分ち書きすることができる。

ただし、複合語の登録については、「単位認定の方針を定めずに、場当たり的に語を登録すると、さまざまな粒度の単位で語が切り出され、単位の不均質性が生じ、語彙調査などの研究で形態素解析ソフトを利用する際に、問題が生じる[3]。」との指摘がある。

本研究の対象である医療情報には、専門用語をはじめ、略語、隠語、外来語、そして造語が含まれる。これらの語種構成や品種構成、造語の規則などの実態は明らかになっておらず、現時点で単位認定を定めることは困難である。

そこで ComeJisyo の登録においては、当面の間は単位認定を規定せず、臨床看護の経験者が「一つのまとまった語」と判断したものを1単位として登録している。

3. ComeJisyo の概要

本章では、2008年11月に作成公開した ComeJisyoV1 から 2011年12月に公開した ComeJisyoV3-1迄の4種の概要を述べる。

3.1 ComeJisyoV1

ComeJisyoV1 は、①Web で公開されている、または研究目的で利用許可を得た看護領域文書に含まれる 50,805 語、②看護学教科書の索引より抽出した 40,833 語、③看護師国家試験問(2002年～2007年)からの抽出語 9,478 語、④Web で公開される用語辞書の登録語 48,875 語の約 15 万語から、複数の出典を持つ用語 30,146 語を登録語とし、半角文字は全て全角文字に変換した上で、NAIST-jdic と統合した Linux 版システム辞書、Windows 版パッケージ、CSV

形式の ComeJisyoV1 の 3 種類を作成し、2008 年 11 月に公開している。

新聞記事 1 年分から作成された機械学習用コーパスによりコストを求めているため、ComeJisyoV1 に登録されているにも関わらず､｢看護｜師｣のように NAIST-jdic0.4.3 に登録されている「看護」が優先し、過分割されるものが多くある [4]。

3.2 ComeJisyoV2

2010 年 1 月に公開の ComeJisyoV2 は、ComeJisyoV1 の 30,146 語に、3 名の臨床管理栄養士が選定した栄養管理、栄養指導に用いる 3,996 語を加え（登録語数 34,142 語）、看護師国家試験および管理栄養士国家試験の過去問題より作成した 7,899 文（146,608 語）の機械学習用コーパスと、新聞コーパス 990,025 語により、コストを求めている。

しかしながら、ComeJisyoV1 に比べて、登録語の過分割は少ないものの、「未｜治療」などの過分割がある [5]。

3.3 ComeJisyoV3

ComeJisyoV3 は、鹿児島大学附属病院提供の看護度分類および患者状態項目 [6]と、国立大学法人 A 病院と財団法人 C 病院の電子医療記録から、臨床看護の経験者 3 名が選定した 7,450 語を加え（登録語数 41,592 語）、2011 年 3 月に公開している。

解析精度を向上させるためには、機械学習用コーパスの整備が不可欠であるが、A 病院および C 病院の電子医療記録には、文字化け、入力ミスや変換ミスが少なくなく、機械学習用コーパスの作成が困難であった。

そこで、Mecab 推奨のシステム辞書である IPADIC2.7.0-20070801 の名詞一般のコストの中央値を参考に、コストを “5,000－文字数×2” と設定し、システム辞書ではなく、Mecab 推奨の IPADIC の形式に合わせたシフト JIS コードのユーザ辞書としてコンパイル版と CSV 形式の 2 種類を公開している。

それでもなお、登録語 41,592 語中 318 語（0.77%）が過分割された。

3.4 ComeJisyoV3-1

ComeJisyo の登録語の殆どは複合名詞であるが、専門用語であることからその細分類である「名詞　一般」か「名詞　サ変接続」かの判断は、医療従事者以外の者には困難であるため、臨床看護経験者の判断に委ねていた。

その結果、同じ単語で終わる複合語の品詞にバラつきが目立つようになった。そこで、**最後に現れる単語（語根）の品詞を語全体の品詞とし**、ComeJisyoV3 の品詞の見直しを行い、助詞を含む用語を削除した。また、318 語の過分割を解消するために IPADIC のコストが 5,000 未満の単語を含む登録語については、その値より低めのコストを設定し（詳細は 4 章で後述する）、登録語数 51,542 語の ComeJisyoV3-1 を 2011 年 12 月に公開している。

4. IPADIC のコストの活用

4.1 ComeJisyo での利用

Mecab 用のユーザ辞書のエントリーフォーマットは CSV ファイルが許す範囲で可変調であり、以下のようになっている [1]。

表層形,**左文脈 ID,右文脈 ID,コスト**,品詞,品詞細分類1, 品詞細分類2, 品詞細分類3,活用形,活用型,原形,フリガナ,ヨミガナ,その他,……

左文脈 ID および右文脈 ID については、“-1” を指定することで品詞情報を基に自動的に mecab-dict-index の実行により付与される。従って、登録語毎に設定が必要なのはコストの値のみとなる。

Mecab の説明書 [1]には、「コストとは、その単語がどれだけ出現しやすいかを示し、小さいほど、 出現しやすいという意味になる。似たような単語と同じスコアを割り振り、 その単位で切り出せない場合は、徐々に小さくしていけばよい。」と記載されているものの具体的な設定方法は分からない。

そこで、ComeJisyo の登録語の多くは名詞であることから、IPADIC の名詞一般のコスト値の分布を調べたところ（表 1）、平均値 5,433、中央値 5,622 であった。

ComeJisyo の登録語は、複合語が多く、IPADIC の登録語に比べて文字長が長い。そこで ComeJisyoV3 のコスト値の設定では、IPADIC の平均値、中央値より小さな値 5,000 を基準とし、文字長×2 を引いた値を設定した。

表１. IPADIC の名詞一般 60,477 語の分布

コスト上限	頻 度	累積割合
-7,000	0	0.00%
-5,600	2	0.00%
-4,200	1	0.00%
-2,800	2	0.01%
-1,400	2	0.01%
0	12	0.03%
1,400	42	0.10%
2,800	255	0.52%
4,200	9,707	16.57%
5,600	8,939	31.35%
7,000	36,401	91.54%
8,400	4,705	99.32%
9,800	277	99.78%
11,200	81	99.92%
12,600	37	99.98%
14,000	9	99.99%
15,400	2	100.00%
16,800	3	100.00%

※ 階級数はスタージェスの公式
（$=1+\log_2$(登録語数)）より求め、階級幅は、
$\frac{\text{(最大値−最小値)}}{\text{階級数}}$の値から切りの良い値を設定

しかし、「救急｜病院」、「急性期｜病院」、「個人｜病院」、「循環器｜病院」、「精神｜病院」など、「病院」で終わる複合語はすべて過分割された。

これは、IPADIC における「病院」のコストが「−3,759」であることに起因していると考えられる。

そこで、ComeJisyoV3 の登録語を IPADIC 辞書で単語分割し、登録語に含まれる単語のコストの最も小さな値を基準にそれより小さな値を登録語のコストとして設定し、改定版 ComeJisyoV3-1 を作成し、公開した。

例：
ComeJisyoV3 におけるコスト
「救急病院」のコスト⇒ 4,982＝5,000−4×2
改定版 ComeJisyoV3-1 におけるコスト
「救急病院」のコスト⇒−4,982
「救急（1,768）」「病院（−3,759）」より

4.1 ６種のコスト実験と考察

ComeJisyoV3−1 の見出し語 41,542 語について、Mecab0.98 を用い、以下のユーザ辞書で単語分割を行った。

(1) ComeJisyoV3-1：
基準値（5,000−文字長×2）の登録語の中に IPADIC のコストの小さな単語が含まれる場合、そのコストより小さい値を登録語のコストとする。
(2) ComeJisyoV3：
コスト値を（5,000−文字長×２）に設定
(3) ComeJisyoV3：
全てのコストを 5,000 に設定
(4) ComeJisyoV3-1000：
全てのコストを 1,000 に設定
(5) ComeJisyoV3-500：
全てのコストを 500 に設定
(6) ComeJisyoV3-0：
全てのコストを 0 に設定

表 2 は、正しく分割されずに過分割された語をまとめたものである。

一律にコスト値を設定する場合、最も効果的なコスト値は 1,000 となった。

しかし、3 語～7 語に過分割されるものが多く、文字数の多い複合語の単位コストとしては適切とは言い難い。文字数が 2、3 文字の単語からなるユーザ辞書の場合、基準のコスト値として 1,000 が適当と考えられるが、医療電子記録に含まれる用語は、いくつもの単語が連結し文字数が多い。

表 2. 正しく分割されず過分割された語数

	全 41,542 語における過分割語数	2 語	3 語	4 語	5 語	6 語	7 語
(1) ComeJisyoV3-1	6	5	1				
(2) ComeJisyoV3	318	273	37	6	2		
(3) ComeJisyoV3-5000	540	493	43	2	2		
(4) ComeJisyoV3-1000	256	135	95	17	7	1	1
(5) ComeJisyoV3-500	812	643	117	39	11	1	1
(6) ComeJisyoV3-0	1,013	694	238	60	19	1	1

実際、ComeJisyoV3-1 の登録語 51,542 語の文字長の最頻値は 4 文字、中央値は 5 文字、平均値は 5.6 文字であり、最大値は 26 文字（コロプラストアシュラコンフォートＬＣ－ＥＣプレカット）である。

そして、IPADIC 内の品詞が「名詞一般」である 60,477 語において、文字数が 3 文字のコストの平均は 490、4 文字では 5,425、5 文字では 10,237 である。

これらのことから、複数の単語からなる複合語については、基準値を 5,000 程度に設定した上で、IPADIC に構成する単語のコストが含まれる場合は、それらの最小値より少ない値を複合語のコスト値とするのが妥当だと考えられる。

5. まとめ

医療用語の標準化がなされないまま、電子カルテシステムは普及し、2009 年時点における全国の 400 床以上の病院の普及率は 41.6%となり、医療機関では電子的情報交換やデータベースの作成が困難になっている。

一方医療情報学の分野では、日々蓄積される医療情報からのテキストマイニングや統計処理に関する研究が行われるようになり†、単語分割には形態素解析器とユーザ辞書が利用される場合も少なくない。

しかし、これらの中には、ユーザ辞書に単語登録することで、正しく単語分割されると誤解している場合や、コストまで考慮せずに独自のユーザ辞書を作成し利用している場合も見受けられる。

これらの医療情報の研究の質を一定レベルに保ち、メタ分析を可能にするためには、単語分割の解析精度を一定に保つことが大切であり、そのためには共有できるユーザ辞書が必要であると考え、ComeJisyo を作成し公開している。

今回、Mecab 用のユーザ辞書のコストについての実験より、2、3 文字の単語登録であれば、コストの基準値は 1,000 程度、また 4 文字以上の複合語については、基準値 5,000 程度とし、IPADIC のコストを参考に調整することで、有る程度の解析精度が得られることが分かった。

ComeJisyo が、医療情報の言語処理において、共有のユーザ辞書として認知され、利用されるようになるための課題としては、以下のものがある。

1） 登録語の単位規則の検討
2） 複合語の品詞規則の検討
3） 登録語の拡充
4） 単語分割、コストに関する調査研究
5） 多義の欧文略語のヨミガナの検討
6） 複数あるヨミガナの検討

謝辞

Mecab の開発者である工藤拓氏、並びに NAIST-jdic の開発メンバーである浅原正幸氏に感謝致します。

本研究は、科学研究補助金　基盤研究（B）「コメディカル実践用語辞書データベースの作成」（課題番号　21300099）の支援を受けています。

参考文献

1） Mecab　http://mecab.sourceforge.net/
2） ComeJisyo
http://sourceforge.jp/projects/comedic/
3） 言語処理学会編：言語処理学事典、共立出版、2009、p.142
4） 相良かおる、浅原正幸、小野正子、小作浩美：形態素解析エンジン Mecab 用看護用語ユーザ辞書の作成と公開、第 28 回医療情報学連合大会、2008、938－939
5） 相良かおる、小野正子、鈴木隆弘、嶋田元、小作浩美：看護記録文の計量的用語調査、人文科学とコンピュータシンポジウム、2010、103－110
6） 鹿児島大学医学部附属病院、看護度分類、
http://www.umin.ac.jp/kagoshima/

連絡先　相良かおる
〒803－0835　北九州市小倉北区井堀 1-3-5
西南女学院大学　保健福祉学部
sagara@seinan-jo.ac.jp

† 2012 年 1 月 23 日時点 Google Scholar で「医療情報」と「マイニング」の And 検索結果は 134 件、「医療情報」「形態素解析」「マイニング」の And 検索結果は 20 件であった。