情報知識学会 第22回(2014年度)年次大会

ComeJisyoの紹介と医療情報に含まれる誤字調査

Introduction of ComeJisyo and investigation of typographical errors contained in medical information

相良かおる¹

Kaoru SAGARA¹

1 西南女学院大学

Seinan Jo Gakuin University

〒803-0835 北九州市小倉北区井堀1-3-5

E-mail:sagara@seinan-jo.ac.jp

電子カルテシステムの普及により、テキスト形式の医療情報が蓄積される。我々はこれら医療情報の自然言語処理を支援するために、形態素解析器MeCab用の実践医療用語辞書ComeJisyoを作成し、現在登録語数77,760語のComeJisyoV5-1を公開している。

また医療施設で蓄積された誤字を含む医療情報を自然言語処理する上で、医療情報にどのような誤字が含まれるのか詳細を知ることが重要である。そこで、倫理的配慮のなされた2施設の医療情報に含まれる誤字の調査を行い、53種の誤字について分析したところ、ローマ字⇒仮名⇒漢字の2段階の変換作業の中で、仮名⇒漢字の変換作業での誤変換が多く、46種が同音異字または同音異義語に誤変換されたものであることがわかった。

The increasingly widespread use of electronic health records systems mean that large amounts of medical information are accumulated in text format. In order to support computer readability and analysis of text by natural language processing (NLP) of medical information, we have released ComeJisyoV5-1 with 77,760 entries of medical terms for morphological analysis of text.

Furthermore, in order to aid NLP of medical text, knowing the kinds of typographical errors present is significant so that they can be reduced and that the terminology can still be interpreted by the computer. Overall, 53 kinds of typographical errors were found and analyzed following an ethically approved investigation of medical information in two facilities.

As a result, in the two-step conversion process, whereby typed Roman alphabetic characters are converted into kana and then exchanged for selected kanji, most errors occurred in the process of exchanging kana for kanji resulting in 46 terms being converted into homophones or another same-sounding kanji.

キーワード: 医療用語, 誤変換, 仮名漢字変換, 形態素解析, 辞書

Key Words: Medical terminology, typographical errors, Kana-kanji conversion, Morphological analysis, dictionary

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。そして現在、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。

そこで、医療情報の分かち書きを目的に、論理的配慮のなされた医療情報より用語を抽出し形態素解析器MeCab^[1]用のユーザ辞書ComeJisyo^[2]を作成し公開している。

電子化された医療情報では、仮名漢字変換システムが介在するため紙カルテの誤字脱字とは異なる人為的なミスが発生しており、実際に医療情報668,995語（ユーザ辞書としてComeJisyoV2を使用）から、ComeJisyo登録語の抽出過程で、誤字だけでなく誤字なのか造語なのか判断の難しい用語も見つかっている。

そこで誤字を含む医療情報の言語処理を支援するために、誤字の調査を行った。

なお、本稿では字形や使い方が誤っている文字だけでなく、誤字を含む単語や複合語も「誤字」と表現している。

本稿では、ComeJisyoの概要と、医療情報に含まれる誤字について述べる。

2 ComeJisyo の概要

2.1 作成方針

一般に単語間にスペース等の区切りのない日本語における単語の単位認定は、形態素

解析辞書に登録された単語によって決まる。すなわち、形態素解析辞書に単語ではなく複合語に登録することで、複合語を単位として分割することが可能となる。

例えば、「時折失見当識見られ、家族不在時にNSコールが押せるか要観察。」という文を一般的な辞書で形態素解析すると、「時折 | 失 | 見当 | 識 | 見 | られ | , | 家族 | 不在 | 時 | に | NS | コール | が | 押せる | か | 要 | 観察 | . 」となるが、「失見当識」、「NSコール」、「要観察」を登録したユーザ辞書を用いることで、「時折 | 失見当識 | 見られ | , | 家族 | 不在 | 時 | に | NS | コール | が | 押せる | か | 要観察 | . 」というように分かち書きすることが可能となる。

なお、形態素解析辞書の登録については、「単位認定の方針を定めずに、場当たりに語を登録すると、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、語彙調査等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。

現実には、医療施設で蓄積される医療情報に含まれる専門用語をはじめ、略語、隠語、外来語、そして合成語の語種構成や品種構成の実態は明らかになっていないため、単位認定を行うのは困難である。そこで、ComeJisyoの登録語においては単位認定の規則を定めず、臨床看護の経験者および管理栄養士が「一まとまりの語」と認識したものを登録語としている^[4]。

2.2 ComeJisyoの概要

① ComeJisyoV1：登録語数30,146語

看護学教科書の索引語、看護師国家試験問題（2002-2007年）に含まれる用語、看護領域の文書より抽出した用語、Web上で公開されている用語辞書の4種の言語資源に共通に出現する用語を選定し、2008年11月に公開。

② ComeJisyoV2：登録語数34,142語

臨床管理栄養士3名により選定した栄養管理・栄養指導分野で使われる用語3,996語を追加し、2010年1月に公開。

③ ComeJisyoV3：登録語数41,592語

倫理的配慮のなされた2施設の医療情報、および看護領域の文書から、臨床看護経験者3名が抽出・選定した用語7,450語を追加し、2011年3月に公開。

④ ComeJisyoV4：登録語数77,647語

新たに1施設を加えた3施設の医療情報より、臨床看護経験者3名が抽出・選定した用語36,055語を追加し、2012年1月に公開。

⑤ ComeJisyoV5：登録語数77,775語

「根性（コンセイ）」、「呼名（コメイ）」、「嫌気（ケンキ）」等、医療分野特有の音読みをする語を含め128語を追加し、2012年3月に公開。

⑥ ComeJisyoV5-1：登録語数77,760語

ComeJisyoV5の見出し語を見直し、更に、教育目的での利用を考慮し、看護師国家試験問題（1998年～2013年）記載の英語訳またはライフサイエンス辞書^[5]の英語訳を付加した。また、看護師国家試験問題、管理栄養士国家試験問題、看護師及び管理栄養士養成大学で使われる教科書の索引語に出現した登録語には、その出所情報を付加し、加えて登録語の出現文書数を付加した。なお、出現文書とは、医療施設3施設の医療情報と言語資源協会より入手した模擬診療録テキスト・データの計4種である。2013年11月公開。

3 誤字の調査

調査手順は以下の通りである。

Step 1. 2施設から提供された倫理的配慮済みの医療情報をComeJisyoV2で分ち書きして得た668,995語よりComeJisyoV3の

登録語を抽出する過程において見つめられた誤字および意図した語ではないと判断した語（総称して「誤字」という）を臨床看護経験者3名により精査し、53種の正誤表を作成。

Step 2. 53種の誤字パターンの分析を行う。

Step 3. Step 1の医療情報とは別の小児看護と小児外科領域の医療情報、A病院：10,346,673文字とB病院：10,615,689文字を対象に正誤表の用語の抽出を行い、出現頻度を求める。

4 結果と考察

4.1 誤字パターン分析

Step 2の結果、誤字53種の内、46種が同音異字または同音異語に誤変換されたものであった。以下にStep 2で分類した9パターンについて述べる。

4.1.1 単純な変換誤り

一文字ずつ入力・変換したことで生じたと思われる単純な変換誤りが11種。その内、「際手」や「件音」は単語単位（「さいしゅ」、「けんおん」）で入力し変換することで避けられる変換誤りであり、この種のものが8種あった。

表1. 単純な変換誤り例

誤	正
T字体	T字帯
回診字	回診時
右腸骨 際 手 部	右腸骨 採取 部
夕方 件 音 時	夕方 検温 時
腓 化 性 嚢胞	腓 仮性 嚢胞
副 用 方法	服用 方法

4.1.2 字形が似ている文字に誤変換

漢字の部首または一部分が同じ、同音の漢字に誤変換された誤字が13種、同音異義語が1種（線種と腺種）あった。これらは意図される意味の類推が容易なため、誤りに気付き

ながら訂正されないものもあると考えられる。

表2. 形状の似ている漢字による誤変換例

誤	正
抹消点滴	末梢点滴
血倍	血培
狭搾音	狭窄音
湿性咳嗽	湿性咳嗽
線種	腺腫

4.1.3 一部の単語の誤変換による異義語

複合語内の一部の単語が同音異義語に変換され、臨床看護の経験者にとっては別の意味を持つ複合語となるものが5種類あった。

表3. 一部の単語が同音異義語に誤変換

不適切	正
安静解除	安静介助
点滴付加	点滴負荷
事故抜管	自己抜管
事故抜去	自己抜去
麻醉下	麻醉科

4.1.4 一部異義語に誤変換された誤字

複合語内の一部の単語が、字形の似ている漢字または同じ漢字を含む同音異義語に誤変換され、複合語としては意味を持たないものが7種あった。

表4. 一部異義語に誤変換された誤字例

誤	正
GDM 母胎児	GDM 母体児
ふらつき著名	ふらつき著明
苦渋様願望	苦渋様顔貌
散在内服	散剤内服

4.1.5 分節区切りの位置の違い

文節の変換区切りの不適切さが原因と思われる誤字が2種見つかった。

表5. 変換対象の違いによる誤変換

誤	正
好酸 急性肺炎	好酸球 性 肺炎
脾臓 脈 塞栓術	脾動脈 塞栓術

4.1.6 意味の誤解によると思われる誤字

「GE(glycerin enema)浣腸」と「サポート支援」は、外来語または略語の意味理解の間違いによる重複表現である。

また、「熱形変化」と「不定形変化」は、「型(type)」と「形(shape)」の意味理解が不十分であることに起因した誤字だと推測され、これらは単語単位の変換により避けられる誤りである。

表7. 意味の誤解によると思われる誤字例

誤	正
GE 浣腸施行	GE 施行
サポート支援	サポート
熱 形 変化	熱型 変化
不定 型 発疹	不定形 発疹

4.1.7 読み方の誤記憶による誤字

漢字の読みの間違いが原因と思われる誤字が3種あり、内2種は字形が似ている。

一方、「声がけ」については、福岡地域では鼻濁音がないため間違いと判断されたが、2施設の所在地は鼻濁音のある地域であり、話し言葉の方言として「声がけ」が使われている可能性がある。

表8. 読み方の誤記憶による誤字

誤	正
症状憎悪	症状増悪（ぞうあく）
粘調痰	粘稠痰（ねんちゅうたん）
声がけ	声掛け、声かけ

4.1.8 イメージが加味されて生じた誤変換

「痰黄色」は、痰の色を表現する際に、「痰」のイメージに影響され、また字形の似た同音異字であることから誤変換されたと推測され、「照灼」はレーザー光線を使って焼灼するイメージから誤変換されたと推測される。

そして「背音」は聴診する部位に、「脳層」は脳の断面画像のイメージに影響を受けた誤変換だと推測される。

表9. イメージが加味されて生じた誤変換

不適正	正
痰黄色	淡黄色
照 灼	焼 灼
背音 Clear	肺音 Clear
脳層ドレーン	脳槽ドレーン

4.1.9 発音の表記の揺れ

外来語の発音の表記の揺れが1種あった。

表10. 発音の表記の揺れ

不適正	正
電動ベット	電動ベッド(bed)

4.2 誤字出現頻度調査

Step 3では、これら53種の誤字の出現状況を調べるために、小児看護および小児外科で蓄積された2施設の医療情報（A病院：看護記録10,346,673文字，B病院：経過記録10,615,689文字）と53種の正誤表に「粘ちゅう」、「ねんちゅう」等の表記の揺れを追加し照合を行い、出現頻度を求めたところ、53種の内の7種（表記の揺れを含めると10種）が正誤表と一致した（表11）。

なお、A病院の看護記録は看護師のみが入力し、B病院の経過記録は看護師と医師が入力しており、語種構成も異なっている^[4]。

表11より、「粘稠」を「粘調」とする誤りが多く、これは「稠」が常用漢字外であること、そして「ちゅう」ではなく「ちょう」と入力した結果の誤りだと分かる。「増」と「憎」は「悪寒（おかん）」、「悪阻（おそ）」等、医療用語には「悪」を「お」と呼ぶ用語があることから、「ぞうあく」を「ぞうお」と入力した結果の誤りだと分かる。またこれら2種の誤字は字形が似ており、意図する意味の類推が容易なために訂正されず、加えて仮名漢字変換システムの自動学習機能も影響し、誤変換されやすく、出現頻度が高くなっていると考えられる。

一方、3名の臨床看護経験者によると、紙カルテでは「誤（抜管、抜去）」を用い、「事故（抜管、抜去）」は使わなかったとのことであった。「誤（抜管、抜去）」の意味には、「自然（抜管、抜去）」、「自己（抜管、抜去）」が含まれる。そして表11において「誤（抜管、抜去）」の意味で「事故（抜管、抜去）」が使われていると思われるものが66語中20語あり、このことから、誤変換により新語が生み出されることが分かった。

表11. 出現頻度調査結果

正	A病院	B病院	① 計	誤	A病院	B病院	② 計	割合 ② / (①+②)
末梢	2,112	263	2,375	抹消	5	0	5	0.2%
血培	84	32	116	血倍	14	0	14	10.8%
自己抜管	3	0	3	事故抜管	9	4	13	80.1%
自己抜去	46	29	75	事故抜去	51	2	53	41.4%
増悪	726	572	1,298	憎悪	46	14	60	4.4%
粘稠	500	34	534	粘調	92	84	176	24.8%
粘ちゅう	22	1	23	粘ちょう	8	0	8	25.8%
ねんちゅう	0	0	0	ねんちょう	17	0	17	100%
声かけ	119	181	300	声がけ	17	13	30	9.1%
声掛け	358	58	416	※ 鼻濁音のある地方の方言の可能性有り				
ベッド	1,250	669	1,919	ベット	53	59	112	5.5%
淡黄色	160	7	167	痰黄色	4	1	5	2.9%

注) A病院の看護記録：10,346,673文字，B病院の経過記録：10,615,689文字

5 まとめと今後の課題

今回、医療情報に含まれる誤字の多くが字形の似た同音異字であること、誤変換が新語を生み出すことが明らかとなった。

誤変換が生じる要因には、人的な入力ミスに加え、仮名漢字変換システムの自動学習機能や予測変換機能、そして、照明や入力画面の大きさ、中腰での入力など人間工学的な環境の問題も含まれる。

手書きカルテと電子カルテの比較において、質的尺度では、紙カルテと電子カルテに複雑さ（自然文らしい記述の多さ）以外の有意な差は認められないものの、量的尺度において、電子カルテの記述量（医療表現および医療用語の数）が多いとする報告がある^[6]。

一方、手書き入力の有効性に関して、①変換キーや訂正キー等の特殊キーの打鍵が多くなり、手書きに比べ入力ミスが多くなることから、手書き入力に比べ、キーボード入力の方がメモ書きの入力速度が遅くなること、②仮名漢字変換により認知的負荷が増し、一時記憶に負担がかかり、思いついた事柄や記憶した内容を忘れてしまい、結果として手書き入力の方が内容の漏れが少なく正確であること、③キーボード入力では、無意識にキーボードに視線が向けられるため、映像および音声情報をメモする際に、大切な情報を見落とす可能性があり、手書き入力の方が映像及び音声情報の理解度が高いとの実験結果が報告されている^[7]。

そして手書き入力インタフェースの研究は、現在の所、手の甲や紙にボールペンで直接メモする程の有効性は期待できない。

すなわち、医療情報を自然言語処理するためには、誤変換についての配慮が必要となる。

そこで今回の調査を踏まえ、誤字であるこ

との識別子を属性として付加した上で、ComeJisyoの登録語に誤字を含めることも検討して行きたい。

謝辞

共同研究者の小野正子氏、上野恵子氏に深謝致します。

本研究は、西南女学院大学共同研究費、並びに、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」の支援（課題番号 21300099）を受けました。

参考文献

- [1] MeCab :
<http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html> （2014年4月4日参照）
- [2] ComeJisyo:
<http://sourceforge.jp/projects/comedic/>
（2014年4月4日参照）
- [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；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，pp.103-110，2010.
- [5] ライフサイエンス辞書プロジェクト：
<http://lsd.pharm.kyoto-u.ac.jp/ja/index.html>
（2014年4月4日参照）
- [6] 荒牧英治；増川佐知子；山田恵美子；脇嘉代；大江和彦：「電子化により外来カルテ文章の情報量は増加する傾向がある」，第30回医療情報学連合大会，pp.1046-1049，2010.
- [7] ムハマド ズルキフリー；田野俊一；岩田満；橋山智訓：「日本語のメモ書き作業における手書き入力の有効性」，電子情報通信学会論文誌，Vol.J91-D，No.3，pp.771-783，2008.