特　集「パーソナルデータに基づく気付きの創発」

患者と医師が使う言葉の違い —闘病記の医学的な応用に向けて—

Difference between Patient Language and Medical Language —Toward a Medical Application Using a Record Written by a Patient—

荒牧　英治　京都大学デザイン学ユニット
Eiji Aramaki　Unit of Design, Kyoto University.
eiji.aramaki@design.kyoto-u.ac.jp, http://mednlp.jp

増川　佐知子　株式会社 Photonic System Solutions
Sachiko Maskawa　Photonic System Solutions.
sachiko.maskawa@gmail.com

宮部　真衣　京都大学デザイン学ユニット
Mai Miyabe　Unit of Design, Kyoto University.
mai.miyabe@gmail.com, http://mednlp.jp/~miyabe

森田　瑞樹　東京大学知の構造化センター
Mizuki Morita　Center for Knowledge Structuring, The University of Tokyo.
morita.mizuki@gmail.com

Keywords: medical informatics, natural language processing, electronic health record, life log, disease journal.

1. はじめに

自然言語から著者の特徴を知る研究が注目されている．例えば，文章分類を応用することで，カルテに記述されている患者が喫煙歴があるかどうか，を一定の精度で判定できる[*1][i2b2 06]．情報抽出を応用することで，副作用の原因となった薬剤を抽出できる [Aramaki 09]．文章検索を応用することで，自分と似た症状の患者を検索することができる[*2]．このような新たな試みはカルテといった病院内文章だけでなく，闘病記やブログといった患者が記述した文章へ適応することで，医療施設ではわからない情報，例えば治療に対する満足度やQOL（生活の質），を抽出できるものとして期待が集まっている．

そもそも，これまで医学的な研究で扱われてきた主なデータは検査値などのラボデータや臨床所見であった．これらは正確であるものの必ずしもすべてのデータが網羅的にデータベース化されているとは限らない．特に，臨床所見に関しては，カルテ中では自然言語で記述される場合もあり，これをデータベース化することは人手が必要となる．

そこで，前述したように，自然言語処理を用いて，臨床データを自動抽出する技術が急速に開発されつつある．例えば，米IBMとMayo Clinicはc TAKES (clinical Text Analysis and Knowledge Extraction System) [Savova 10][*3]という自由記載の臨床記録から情報抽出するシステムを研究開発している．コロンビア大学のMedLEE [Sevenster 12]は放射線読影レポートについて情報抽出を行う．日本語についても，東京大学と富士ゼロックスがTEXT2TABLEという退院サマリ用の情報抽出システムを開発している [Aramaki 09]．また，情報抽出の精度を競うコンテスト形式のワークショップ（シェアードタスク）も活発に行われている[*4][i2b2 06]．同様のワークショップは，本邦でもNTCIR MedNLP[*5]として本年度から開催されている [Morita 13]．このように多くの試みがなされているが，これらはすべて医療者の作成したテキストを対象としたものであり，データの使用や配布について制限がある．

このような状況の中，近年，新たな試みとして，患者が自ら発する文章から臨床情報を抽出する試みが注目を集めはじめている．患者の記述する文章はブログ，ソーシャルメディアからメールまで膨大な種類と量があ

*1 https://www.i2b2.org/NLP/
*2 http://www.naika.or.jp/info/rireki/info110617.html
*3 http://ctakes.apache.org/
*4 https://www.i2b2.org/NLP/
*5 http://mednlp.jp/medistj-ja/

り，医療者の文章量をはるかに凌駕する．これらの医学的な応用を試みた研究としては，検索クエリ [Ginsberg 09, Polgreen 08] やソーシャルメディア上での発言 [Aramaki 11, Paul 11] を材料に感染症の流行を推定する研究や，ソーシャルネットワーキングサービス（SNS）として疾患情報を共有・収集する PatientLikeMe *6 などがある．これらは，患者自身の力で治療をサポートする Self-Management の一種とも考えられ，今後ますます発展すると予想される [Ellis 13]．これら患者テキストを用いたサービスの利点は2点ある．

(1) **量の膨大さ**：患者自身がデータを記述するため膨大なデータ量が得られる．

(2) **質的な新しさ**：従来，看護記録に記録されるにとどまっていた精神面や QOL に関する記述が豊富に得られる．

これらの性質は，大規模な調査が困難な難病や，対応の迅速性が求められる感染症などと相性が良く，今後の応用が期待されている．その一方で，医学的な専門知識のない患者が記述するため信頼性の問題がある．例えば，何による痛みか勘違いして記述したり，薬を飲み忘れていたのを隠して飲んだように記述することも考えられる（**信頼性の問題**）．

さらに，問題となり得るのは，いくら正確に記述したとしても，医学的な用語との乖離が大きく，医学的な利用が困難な場合がある（**用語の差異の問題**）．例えば，「指先がピリピリ痺れています」という記述は，副作用報告としては「末梢神経障害」と言い換えて集計する必要がある．

ここで前者の**信頼性の問題**は真の症状との対比調査が必要となりその実態調査は困難であるが，後者の**用語の差異の問題**は患者文章を大量に収集することで検討が可能である．患者が用いる言葉はどのように医学用語から逸脱するのであろうか？

2. 材　　料

2·1 闘病記とは

一般の検索エンジンにて「闘病記」を検索した場合，闘病記専門の検索サイト，書籍の闘病記などが該当し，闘病記そのものを得ることは難しい．また，そもそも，どのような文章が闘病記であるかも明確な定義がなされていない．そこで，本稿では，以下の二つの闘病記専門の検索サイトを用い，得られたテキストを闘病記とみなした．

(1) TOBYO（株式会社イニシアティブ）*7：疾患名による検索を行うと，闘病記が掲載されたブログについて，ブログ運営会社を問わずピックアップすることができる．さらに，患者の性別，ブログ開設時期，ブログの評価順などから絞り込むことも可能である．

(2) LifePalette（株式会社メディエイド）*8：闘病に関する患者ソーシャルネットワーキングサービス（SNS）．疾病に関わる情報掲載や，執筆者同士の情報交換が可能となっている．

表1 闘病記の例（胃がん（上）と認知症（下））

化学療法，始まりました ２０１１年2月２５日 ついに抗癌剤が２月２０日に投与されました． 本当は１５日に始める予定だったのですが，投与前の血液検査で白血球の数値が低かったため，延期となっていたのです． 後日の血液検査では問題なかったため，２０日スタートです． お薬の名前は「タキソール」というもの． そして心配していた副作用が，やはり出ています． まず発熱，３８度近く出ました． あとは関節痛や筋肉痛があり，とにかく全身倦怠感． だいたい予想していた症状が出た感じです． それから昨日あたりから，指先がピリピリ痺れています． びっくりして看護師さんに聞いてみると，これも副作用のひとつだそうです． いろいろあるんですね …． 何だか憂うつですが，頑張って耐えようと思います．
２０１３年3月４日 タイトル：おばあちゃん，怒る ２０１２年１２月頃の話です． 久々に祖母から電話がありました． しかし開口一番，すごい剣幕で怒っていました． 「この間のことなんだけどね，私はあんなこと言ってない!!」 話の内容を聞いたのですが，どう考えても実際にはなかった出来事について怒っているのです． もしかして，妄想が起きているのでは … 不安で少し怖くなりました．

＊本文は医療従事者によって作成された架空の闘病記録からの抜粋である．

上記サイトから得られる闘病記は，誰が執筆したかという観点から次の二つに分類できる．

(1) 本人執筆型：患者本人が自分の体験記を作成したもの．例を表1上に示す．

(2) 介護人執筆型：患者の家族など，周囲の人が患者の闘病生活を見てつづっているもの．例を表1下に示す．

割合としては前者のほうが多いが，認知症などある種の疾病によっては本人による記載が困難な場合があり，この割合は疾病に依存する．

また，執筆スタイルからも次の二つに分類できる．

*6 http://www.patientslikeme.com/
*7 http://www.tobyo.jp
*8 http://lifepalette.jp

(1) 単テーマ：疾病のみをテーマにしたもの．

(2) 多テーマ：内容ごとにカテゴリーが複数作成されており，その中の一つが闘病記となっているもの．例えば，テーマ「病気のこと」などとして分類されている．

2・2　闘病記の抽出方法

日本人の死因の主要な疾病から，闘病が長く続き闘病記が得られやすいがん，認知症，うつを対象とした．ここで先の闘病記の検索サイトにて検索を行い，上位に表示されたものから無作為に収集した．この結果，計167の闘病記が得られた（表2）．

次に，各闘病記から，罹患時から5記事を無作為に抽出した．ただし，多テーマの闘病記については，疾病と無関係な記事は除いた．この結果，835記事(＝167×5)，24 715文が得られた．

表2　対象とした疾患

疾患	(疾患のICDコード*)	サイト数	記事数
胃がん	(C1)	40	200
肝臓がん	(C2)	10	50
肺がん	(C3)	25	125
大腸がん	(C18)	32	160
認知症	(F0)	20	100
うつ病	(F3)	30	150
乳がん	(C5)	10	50

*ICD（International Statistical Classification of Diseases and Related Health Problems；疾病および関連保健問題の国際統計分類）とは，世界保健機関（WHO）が作成した死亡や疾病のデータの分類．

2・3　用語の抽出方法

抽出した文に対して2名の医療従事者（検査技師と治験レポート集計者）が人手でこれを精査し，患者の症状が読み取れる箇所についてマークアップ（以降，アノテーションと呼ぶ）を行った．アノテーションは2名が分担して（重複なく）行った．

アノテーションの例を表1に太字で示す．太字となっている部分について，さらに，副作用報告レポートとしてどのような用語が適切か，MedDRA/J（Medical Dictionary for Regulatory Activities Terminology）[*9]での分類を行った．この結果，8 943表現がアノテーション・分類された．

3. 結　　果

どのような症状が闘病記に記録されているかを表3に示す．表に示されるように，「不安」，「疼痛」，「悪心」など，患者のクオリティオブライフ（以降，QOL）に関する症状が多い．これらは本来，患者自体が訴えないと記録されないため，患者自身が一次情報として記録することは自然である．

表3　闘病記に出現した症状用語のPT別の頻度と割合

用語	出現頻度	割合〔%〕
不安	122	3.75
疼痛	108	3.32
悪心	83	2.55
倦怠感	54	1.66
下痢	51	1.56
発熱	46	1.41
疲労	37	1.13
貧血	32	0.98
出血	29	0.89
嘔吐	29	0.89
感覚鈍麻	28	0.86
腹水	28	0.86
便秘	28	0.86
不快気分	27	0.83
緊張	24	0.73
浮動性めまい	22	0.67
腹痛	22	0.67
咳嗽	21	0.64
動悸	21	0.64
不眠症	21	0.64
無力症	20	0.61
異常感	19	0.58
処置による疼痛	18	0.55
ストレス	16	0.49

*集計は，人手で基本語に変換した後の集計である．胃がん，認知症など，闘病記の主たる疾患そのものはこの集計からは除いた．

次に，これらの用語の二次利用の容易さという観点から，以下の四つのカテゴリーに分類した．これらと，闘病記でなく医師が記述した文章（以降，医師文章）との比較を行った（表5）．医師文章はNTCI R10 MedNLPタスク [Morita 13] で用いられた模擬病歴報告を材料とした．

(1) **基本医学用語**：副作用レポートで用いられるMedDRA/Jの基本語がそのまま用いられる場合．

(2) **標準医学用語**：基本語とは異なるもののMedDRA/Jに収載されている用語が用いられる場合．

(3) **非標準医学用語**：基本医学用語や標準医学用語以外の用語が用いられる場合．

(4) **症状文**：名詞として表現されず，句や文として症状が表現される場合，例えば「食欲不振」が「少ししか食べられない感じです」と表現され得る場合．

結果，基本医学用語の割合が闘病記39.7%に比べ，医師文章30.0%となり，闘病記が上回った．また，標準医学用語においても，闘病記19.6%に比べ，医師文章16.0%となり，闘病記が上回った．また，非標準医学用語の割合は，闘病記も医師文章40%以上であった．このことは，医療文章といっても必ずしも二次利用に適した用語が用いられているわけではなく，闘病記と医学文章のいずれも，一定の割合で用語を変換する必要があることを示している．

ただし，非標準医学用語の在り方は両者で大きく異な

*9 https://www.pmrj.jp/jmo/

る．医師文章においては，表4に示されるように，「高K血症」といった略語や「両下腿浮腫」といった詳細な表現が多く用いられ，MedDRA/Jの用語から逸脱している．

一方，闘病記における非標準医学用語は表6に示される「ダルさ」や「ヒリヒリ」など日常的な用語が用いられている．これらが医学用語から逸脱する理由は以下のように考えられる．

表4 医師文章に含まれる非標準医学用語

非基本語	頻度
意識清明	12
両下腿浮腫	10
腸音異常	10
腹部平坦，軟	10
肥厚	9
腸蠕動音異常	8
眼瞼眼球結膜異常	7
スリガラス影	6
炎症反応高値	6
平坦，軟	5
易出血性	3
高K血症	3

(1) **難解であるケース**：難解な語は闘病記に記述されない．例えば，「悪性新生物」は闘病記では，「ガン」のようにカタカナ表記される．基本語「疼痛」，「悪心」なども同様のケースである．

(2) **羞恥心を喚起するケース**：羞恥心を喚起し，ネットで公開する文章に記述するのがはばかられる場合は，言換えが行われる．例えば，「下痢」は「下り気味」「ピー」などと柔らかい表現に変換されている．また，「嘔吐」は「リバース」などと言い換えられている．

(3) **擬態語・擬音語などで表現するケース**：「感覚鈍麻」は「ビリビリ」や「ピリピリ」など擬態語で表現されている．また，「倦怠感」は「ダルさ」と表現されている．

今後，闘病記を扱う際には，これらの変換辞書が必要となるであろう．本稿で扱った材料に出現した表現に関しての変換結果はサイト[*10]にて公開している．

最後に，用語の種類（異なり/タイプ）は闘病記773タイプに対し，医師文章1413タイプとはるかに医師文章が多い．これは，闘病記には，同じ表現が複数回出現しており，逆に医師文章は，同じことを記述しない性質によるものと考えられる．これらの性質を考慮すると，網羅性，正確性については医師文章に頼らざるを得ず，

表5 闘病記と病歴報告における症状表現の分類

分類	闘病記	医師文章（模擬病歴報告）
症状句・文リソース	9.0% (=2247文/24715文)	2.4% (=83文/3366文)
基本医学用語	39.7% (=1299語/3269語)	30.0% (=799語/2655語)
標準医学用語	19.6% (=643語/3269語)	16.0% (=426語/2655語)
非標準医学用語	40.6% (=1327語/3269語)	53.8% (=1430語/2655語)
用語の種類（異なり/タイプ）	773タイプ	1413タイプ

*非リソースは文単位，その他のリソースは語単位で集計した．

表6 闘病記に出現した表現とその基本語

基本医学用語	闘病記に出現した表現 (出現数)
悪性新生物	がん (102), ガン (58), がん細胞 (9), 癌細胞 (6), 末期がん (6)
うつ病	鬱 (21), うつ (19), 鬱病 (2), 介護うつ (2), 鬱状態 (2), 抑うつ状態 (2)
胃癌	胃がん (54), 胃ガン (7), 進行胃がん (2), 早期胃ガン (2)
不安	不安感 (8), 夕暮れ時症候群 (1), 不快感 (1), 不安状態 (1)
認知症	認知症状 (3), 痴呆 (2), 認知 (1)
疼痛	痛み (82), 痛い (10), 激痛 (10), 痛さ (2)
悪心	吐き気 (53), ムカつき (4), ムカムカ (4), 気持ち悪さ (4), 吐気 (3)
倦怠感	ダルさ (8), だるさ (6), だるい (1), 億劫さ (1), 倦怠 (1), けだるさ (1)
下痢	ぴー (2), ゲーリーさん (2), 下り気味 (1), ピー (1), 軟便 (1), ゲーリーP (1)
嘔吐	おうど (1), リバース (1)
感覚鈍麻	ビリビリ (8), 痺れ (8), しびれ (7), ヒリヒリ (1), ピリピリ (1)
浮動性めまい	眩暈 (4), 目まい (3), 目眩 (2), ふらつき (1)
咳嗽	せき (2)
不眠症	不眠 (12), 絶不眠 (3), 眠れない (1), 不眠症状 (1), 不眠兆候 (1), 不眠悪化 (1)
無力症	体調不良 (9), 体調低下 (1), 全身衰弱 (1), 衰え (1), ぐったり (1), ダルさ (1), 脱力 (1)
健忘	物忘れ (11)

*10 http://mednlp.jp/resources/

QOLに関する症状について，その程度・頻度などの情報は闘病記に利用可能性があると考えられる．

4. ま と め

近年，患者の記述し得る文章はブログ，ソーシャルメディアからメールまで膨大な種類と量になり，医療者の文章量をはるかに凌駕している．これを有効に利用できれば，大規模調査が困難であった希少疾患や，対応の迅速性が求められる感染症などへの応用が期待できる．その一方で，医学的な専門知識のない患者が記述することによる不正確性も危惧されている．本稿では，患者が執筆した167サイトの闘病記を調査し，医師の記述した文章（病歴報告）と比較した．調査の結果，闘病記に含まれる症状表現の量は不安，倦怠感，不眠症などのQOLに関する情報が多く含まれていた．今後，これをフルに活用するために，逸脱した表現を変換するリソースの開発が必須である．

謝 辞

本稿に扱ったデータの収集にはJST戦略的創造研究推進事業さきがけ「情報環境と人」領域「自然言語処理による診断支援技術の開発」，科学研究費補助金若手研究A「表記ゆれおよびそれに類する現象の包括的言語処理に関する研究」のサポートを受けた．

◇ 参 考 文 献 ◇

[Aramaki 09] Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Mashuichi, H. and Ohe, K.: TEXT2TABLE: Medical text summarization system based on named entity recognition and modality identification, *Proc. Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL2003) Workshop on BioNLP*, pp. 185-192 (2009)

[Aramaki 11] Aramaki, E., Maskawa, S. and Morita, M.: Twitter catches the flu: Detecting influenza epidemics using twitter, *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2011)

[Ellis 13] Ellis, L., Showell, C. and Turner, P.: Social media and patient self-management: Not all sites are created equal, *Studies in Health Technology and Informatics*, Vol. 183, pp. 291-295 (2013)

[Ginsberg 09] Ginsberg, J., Mohebbi, M. H., Patel, R., Brammer, L., Smolinski, M. S. and Brilliant, L.: Detecting influenza epidemics using search engine query data, *Nature*, Vol. 457, pp. 1012-1014 (2009)

[i2b2 06] i2b2: Informatics for Integrating Biology and the Bedside (2006)

[Morita 13] Morita, M., Yoshinobu, K., Tomoko, O., Mai, M. and Aramaki, E.: Overview of the NTCIR-10 MedNLP task, *Proc. NTCIR 10 Workshop* (2013)

[Paul 11] Paul, M. and Dredze, M.: You are what you tweet: Analyzing twitter for public health, *Proc. 5th Int. AAAI Conf. on Weblogs and Social Media (ICWSM)* (2011)

[Polgreen 08] Polgreen, P., Chen, Y., Pennock, D. and Nelson, F.: Using internet searches for influenza surveillance, *Clinical Infectious Diseases*, Vol. 47, No. 11, pp. 1443-1448 (2008)

[Savova 10] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper.-Schuler., K. C. and Chutp, C. G.,: Mayo clinical text analysis and knowledge extraction system (cTAKES) : Architecture, component evaluation and applications, *J. Am. Med. Inform. Assoc.*, Vol. 17, No. 5, pp. 507-513 (2010)

[Sevenster 12] Sevenster, M., Ommering, van R. and Qian, Y.: Automatically correlating clinical findings and body locations in radiology reports using MedLEE, *J. Digit Imaging*, Vol. 25, No. 2, pp. 240-249 (2012)

2013年9月1日　受理

著 者 紹 介

荒牧　英治

2000年京都大学総合人間学部卒業．2002年同大学院情報学研究科修士課程修了．2005年東京大学大学院情報理工系研究科博士課程修了．博士（情報理工学）．2008年東京大学知の構造化センター特任講師．現在，京都大学デザイン学ユニット特定准教授，科学技術振興機構さきがけ研究員(兼任)．医療情報学，自然言語処理応用の研究に従事．

増川　佐知子

1989年お茶の水女子大学理学部物理学科卒業．1991年同大学院理学研究科物理学専攻修士課程修了．1991年花王株式会社入社，数理科学研究所配属．1999年国立福山病院附属看護学校，福山平成大学，福山市立女子短期大学非常勤講師．2010年東京大学知の構造化センター学術支援専門職員．現在，株式会社 Photonic System Solutions 研究員．

宮部　真衣

2006年和歌山大学システム工学部デザイン情報学科中退．2008年同大学院システム工学研究科システム工学専攻博士前期課程修了．2011年同専攻博士後期課程修了．博士（工学）．現在，京都大学デザイン学ユニット特定研究員．多言語間コミュニケーション支援，マイクロブログ上の流言拡散防止に関する研究に従事．

森田　瑞樹

2003年東京工業大学生命理工学部卒業．2005年同大学院生命理工学研究科修士課程修了．2008年東京大学大学院農学生命科学研究科博士課程修了．同年東京大学大学院農学生命科学研究科特任助教．2009年医薬基盤研究所特任研究員．2012年東京大学知の構造化センター特任研究員，現在に至る．生命情報科学，医療分野における自然言語処理の研究に従事．