Web 上のがん情報取得のためのがん用語辞書の作成

木村 俊也† 中川 晋一†‡* 三角 真‡ 山岡 克式* 酒井 善則* 島津 明†

†北陸先端科学技術大学院大学情報科学研究科 〒923-1211 石川県能美市旭台 1-1

‡情報通信研究機構 〒184-8795 東京都小金井市貫井北町 4-2-1

*東京工業大学大学院理工学研究科 〒152-8500 東京都目黒区大岡山 2-12-1

E-mail: †{s-kimura, shimazu}@jaist.ac.jp, ‡{snakagaw, misumi}@nict.go.jp

*{nakagawa, yamaoka, ys}@net.ss.titech.ac.jp

1. はじめに

治療法の確立されていない疾患であるがん[1]を発病したがん患者にとって，日々刻々更新される最新のがん情報を的確に得ることは延命や治癒のために,手術,内服薬に匹敵する第三の薬である．医療現場で医師から与えられる情報の正確さと根拠について，より的確な情報を早期に得，より的確な医療を受けることが予後（生存期間）の延長や余命の中での生活の質向上に直結するからである．

インターネットによるがんに関する情報発信が活発化し，情報の量が増加してきているのは望ましいが,調査により以下のことが判明した.中川・木村[1][2]は，胃がん，肺がん，大腸がん，子宮がん，白血病の５つのがんについて，わが国で発信されているこの分野のコンテンツは１：専門医療機関や教育機関による研究業績などの高度な内容,２：個人医師や患者個人による患者指向の内容,３：個人を対象としたポータルサイトや書籍の情報，４：個人を対象とした商用情報，５：検索ノイズ，の５類型に分類されることを報告した．このうち専門性の高い研究指向の類型１は根拠があり有用な情報を含むが，専門用語の知識のない患者にとって理解することが困難であり，間違った解釈を生むことも考えられる．さらに専門機関による Web 情報発信量は予算によって左右されることも問題である．むしろ類型２の個人を対象とした個人（情報ボランティア）による情報発信からの情報がより患者のニーズに近い情報を与える可能性があることが示唆された.従って類型２(情報ボランティアによる情報発信）を類似ページの抽出や blog でのトラックバック時の候補の抽出時に提示することなどにより，いわゆる患者コミュニティ形成が促進される可能性があると考えられた．そこで今回国立がんセンター(NCC-CIS)の Web データを元に提供されているすべてのがん（計 54 種類）について用語辞書を作成し，がん情報を必要とする患者のために「がん」に関する文章で用いられる言語的特徴を明らかにすることを目的として検討した．

[1] 専門家では、「癌」は固形癌を表す場合が多く、白血病や肉腫などの疾患群を含めるために、国立がんセンターでは敢えて「がん」とひらがな表記する．本研究でもこれを採用する．

2. がん言語空間の特異性と患者コミュニティ形成の必要性

「がん」は他の傷病とは異なり，臨床上極めて特異的な疾患群である．臨床医学関係者であってもがんに対する理解は必ずしも一定ではなく，臓器別，特徴別にそれぞれの種類のがん患者と専門家が存在する．さらに有効な治療法のない場合などに対してはがん性疼痛の軽減を目的としたターミナルケアが臨床上重要であり複雑である．以下,臨床的見地から見たがんの特殊性について述べ,患者コミュニティ形成の必要性と本研究の意義に関して述べる．

2.1. 臨床的特徴

「がん」は，人体を構成する自らの細胞が「がん化」し通常細胞とは異なった性状（増殖，転移）を示すことにより，正常な組織を破壊，閉塞，栄養欠乏などの障害により死に至らせる疾患である.人体を構成する細胞は,もともと内胚葉,中胚葉,外胚葉の３種分かれ，それぞれの器官（代表的なものとして血液，筋肉・内臓，神経）の形成が行われる．原則的にすべての細胞ががん化する可能性があると考えられている．この３つの系統ごとにそれぞれがんの性状は異なる．逆に言うとこの３つの系統内のがんは生物学的性質が類似するため，特に細胞の生理を利用してがん細胞を死滅させるための治療法は共通する場合が多い.例えば,内胚葉由来の骨髄細胞ががん化する白血病と同じ内胚葉由来のリンパ節の細胞ががん化する悪性リンパ腫では，細胞分裂の速度が速く急激に全身に転移するという性質が類似しており，手術よりも抗がん剤を用いた化学療法が治療の中心と成る．また，大腸がんと胃がんは同系統の消化器上皮から生じる腺がん(Aden carcinoma)であり,固形がんである．この種のがんは増殖速度がそれ程速くないことと，高分化細胞由来であるため一般に化学療法は奏効しないことが特徴である．従って予後を左右するのは早期発見と外科的切除である．

さらに手術を行う前のステージング（例えば胃のどこにがんがあるのか，局所的にはどの程度浸潤しているのか，胃の壁を食い破って腹腔に出ているのか，転移はあるのか，あればリンパ節の転移はどこまで広がっているのか，血管を介してど

れ位遠くの臓器まで転移しているのか，などのパラメータを用いた進達度と病期の分類）と，がんの性状（ポリープ状，潰瘍を形成するもの，下掘れ潰瘍を形成するもの,びまん性に浸潤するもの）によって治療方針の選択が異なる．患者はステージングによって治療法も予後（余命の推定値）も異なるため，それぞれ必要とする情報が異なる．

2.2. Web を介した患者コミュニティ形成の必要性

以上のように,「がん」と言っても，患者それぞれが別のがんを発症し，ステージ，宣告された予後が異なる．さらに確かにわが国の死亡率第一位はがん（悪性新生物）であるが，最も数の多いとされる胃がんでも人口１０万人あたり罹患率（発症する人数）は年間男性約９０人，女性３３人，年齢調整死亡率は男性約３５人，女性１５人である[3][4]. また，胃がんの場合健康診断で見つかるステージ別には 1995 年から 1999 年のデータで限局性（転移のないもの）21001, 領域がん（軽度の転移）11660,遠隔（他の臓器への転移のあるもの）5894, 総数 43409 人（調査数）であり，５年生存率はそれぞれ 95%, 40%, 3%, 58%と報告されており，限局：領域：遠隔の比は約 4：2：1 である．従って，胃がんの場合，男女合わせて人口 10 万人あたり男女計約 120 人が罹患し,約 50 名が死亡する．早期胃がんの場合ほぼ完全に治癒するため仮に情報提供を必要とするグループを領域がんとすれば，約 30%すなわち生存者 70 名のうち 3 割（20 名程度）が領域がんの治療法に関して情報を必要とする可能性があると算出できる．つまり人口 10 万人あたり 20 名の潜在的情報要求があると仮定できる．人口 10 万人あたりの病院の設置数は 1 から 2 施設であるから 1 施設あたり約 10 名前後が情報要求を持つと考えられる．この規模の人数に対して専任のがん情報を供与できるスタッフを雇用することは不可能である．また，地域においても 100 万人口の都市でさえ年間 100 名から 200 名の患者を対象とすることになる．この場合，相談室を作っても 1 日 1 名から 5 名程度の相談者となることは容易に予想される．従って本事象は公的扶助の対象として成立させることが困難である．むしろ電子メイルなどのインターネットメディアを活用した情報交換，情報提供の良い適応となると考えられる．

2.3. NCC-CIS の言葉集合の特性

がんの種類は多岐に渡るが,「診断方法（レントゲン写真，CT 検査，MRI），治療方法（外科的切除，化学療法，緩和ケア）」等の語は共通する．解析に使用する標準語群として，わが国で標準的に使われている NCC-CIS で提供されている Web ページをそれぞれのがんについて取得し単語を切り出した．その結果，Leukemia，LC, SC, CC, UC の 5 種類の各がんについての説明文書に出現する専門用語数は各 5 種の文章においてどれも約 300 語であった．それぞれのがんの上位 15 語(５つのがんで合計 75 語), 30(150), 40(200), 50(250), 75(375), 100(500)に出現した単語集合の和集合を x 軸（重複を含む）とし，異なり数の個数を y 軸としてプロットした結果を Fig.1 に示す．各がんにおける語の出現頻度は選択した語数によらず直線的に相関しており，この言葉集合（それぞれのがんにおける NCC-CIS での単語）が一般の URL においてほぼ当確率で出現していることが示された．これら NCC-CIS に用いられている言葉集合に関して国立がんセンターに問い合わせたところ，専任の Peer Reviewer が各コンテンツ作成者の作成した文章に対して一般人の理解が得られやすくするために語の書き換えを指示しており，できるだけ一定の単語で記述するようにバイアスをかけている結果であることのことであった.以上のことから，患者への情報提供の適正化を目的とする本研究において最適と判断した．

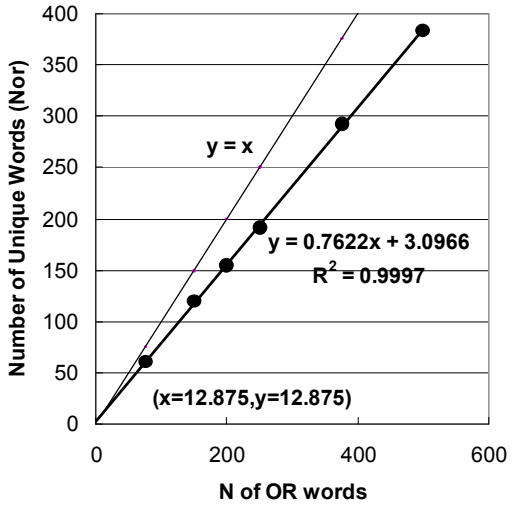

Fig. 1: Result of ‘Unique’ for the simple ‘or’ of Various Rank of 5 Cancers

3. がん用語辞書の作成

国立がんセンターの Web ページの文章を元に用語辞書を作成した．以降，作成の手順を示す．がん専門用語は，がんの国立専門研究機関である国立がんセンター(NCC-CIS)で提供されている疾患別解説ページにそれぞれ出現する単語を切り出した.がんを解説している疾患数は計 54 種類あり，それぞれ手作業で専門用語を切りだした．これらを一つの集合とし，疾患ごとに独立して集合を作成した．こうして本集合の各要素の異なり語を用語辞書として固定した．

4. 妥当性の検討

辞書との妥当性を検討するため，疾患別に用語を加えたときの辞書内に存在する用語数について検討した結果を Fig.2 に示す． 横軸はそれぞれがんの疾患であり，縦軸は得られた専門用語の合計数である．

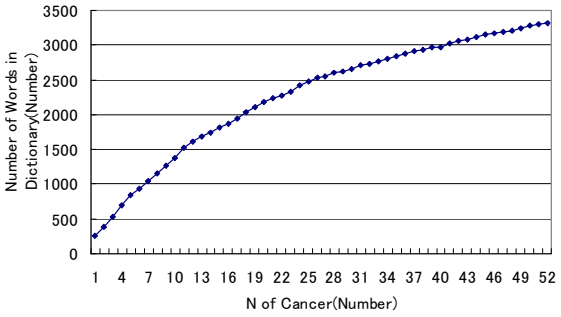

Fig.2 N of words in Dictionary

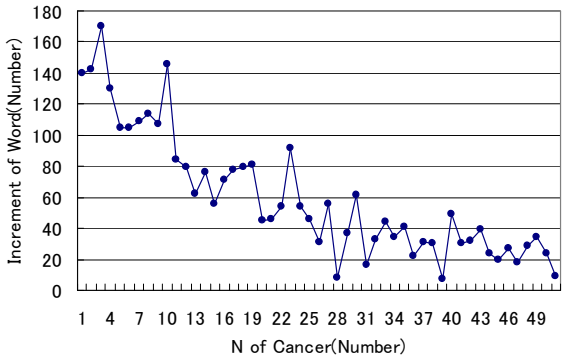

Fig.3 N of increment of words in Dictionary

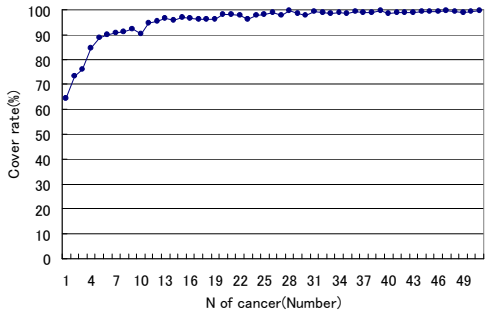

Fig.4 N of cover rate of word using cancer

がんの数を増加させてゆくとともに辞書の単語数も単調に増加するが，1つのがんあたりの増分が減少する．計54種類を合わせた結果，辞書に取り入れる用語は合計3316語となった．Fig.2のがん毎の微分値のプロットをFig.3に示す．

増減があるものの単調減少であり，約10個の疾患で全体の単語数の約25%を約20個で約50%を占める．次に，疾患ごとに用語を加えていく過程で，疾患を一つ加えるごとに，どれほどの用語が重複しているかを示したものをFig.4に示す．横軸には各疾患を，縦軸には一つ疾患を加えたときの重複率を示した．Fig.4に示したように，各疾患を解説するのに用いられる専門用語は多くが重複していることがわかる．以上の考察から我々は「WWW上でよく用いられるがん専門用語は限定されており，標準的な研究機関である国立がんセンターのWebページで用いられている専門用語をがん専門用語辞書に収めれば，大概の専門用語はカバーできる．」という仮説を立てた．この仮説を元に本稿で作成したがん専門用語辞書を用いてどれだけカバーできているか実験を試みた．

5. 実験

作成した専門用語辞書をchasen（chasen-2.3.3 + ipadic-2.7.0)[5]に適用して実験した．実験方法はがん患者，完治済みのがん患者が作成した闘病記を綴ったblogページをテストデータとした．まず，栃木がんセンターのWebページにある，計15種類の臓器別診療情報の文章を形態素解析するのに本研究で作成したがん用語辞書をchasenに適用した結果得られた解析結果と適用しない場合での結果をTable.1に示す．次に，各blogページに出現する専門用語を手作業で分割し，本稿で作成したがん専門用語辞書がどれほどカバーしているかを計測する．まず，検索エンジンであるgoo[6]を用いて，検索語を「がん闘病記」として与えた結果得られたblogページをランダムで30ページ選出した．そしてその30ページに出現する専門用語を手作業で選出した．なお，得られた用語でipadicの辞書に含まれる用語はあらかじめ削除した．その結果各blogページに出現した専門用語数の推移をFig.5に示す．なお，Fig.5の横軸はがん専門用語の出現回数が多いblogページ順に並べた．がんに関する個人が作成したblogページに現れるがん専門用語は平均4.56回と少ないことが示唆された．そして，個々のblogページに出現した専門用語を我々が作成したがん専門用語集がどれほどカバーしているかを調べた結果をFig.6に示す．平均65.1%の用語が辞書にある用語と重複していた．がん専門用語辞書に含まれていなかった用語の一例を，カテゴリに分類してTable.2に示す．

Table.1 result of morphologic analysis with chasen and cancer dictionary

	形態素数	未知語検出数	未知語率(%)
用語辞書あり	25098	134	0.533907084
用語辞書無し	26802	265	0.988732184

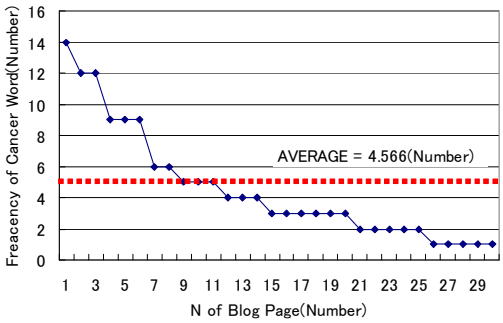

Fig.5 Frequency of words about cancers on blog

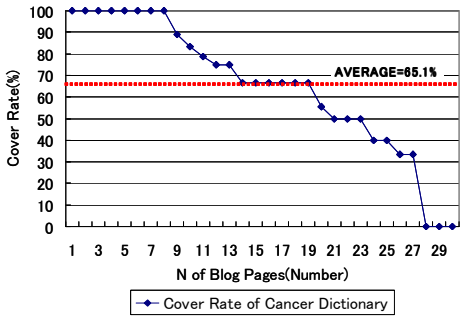

Fig6. Cover rate of cancer dictionary

Table.2 Words not included in dictionary

1群:「がん」の表記のずれ		
抗ガン剤	子宮頸ガン	ガン細胞
2群:薬品名		
アレビアチン	グリオブラストーマ	ジフルカン
ハルシオン	ボルタレン	レドニン
3群:複合語		
MRI画像	完全麻酔	手術前投薬
麻酔前投薬		
4群:治療法		
AdVP療法		

まず 1 群に現れた,「がん」の表記のずれに関して，我々はひらがなで「がん」として表記している．しかし，がんは漢字でもカタカナでも表記できる．がん専門用語辞書に，漢字で「癌」，カタカナで「ガン」を追加すると登録する用語の量が大幅に増加してしまう．これに関しては今後の検討課題にするが，がんに関する情報に対し言語的な何らかの処理をする場合は，得られた情報を一度我々が使用する言語の形式（例えば「ガン，癌」ならば「がん」にする.）に変換してから処理するといった方法を考えている．2 群の薬品に関して，薬品は種類が多く，かつ，新薬が作成される頻度も高い．よって，すべてのものを登録するわけではなく，WWW でよく使われるものの中から，危険性が低く認可されているもののみを登録する方針で考えている．これは 4 群の技術に関しても同様である．3 群に含まれる複合語に関しては依然検討中である．がん専門用語には複合語が多く存在している．表にも示したように，例えば「MRI 画像」という用語がある．我々が作成した辞書には「MRI」と「MRI 検査」が登録されているので，「MRI 画像」が未知語となることはない. しかし，複合語で成り立っている専門用語をすべて一つの形態素とするかを決定しなければならない．本稿では，一部の例外を除いて複合語を一つの形態素として登録した．例外とは，がん専門用語で特有に用いられる「原発性胃がん」や「転移性肺がん」といった「原発性」や「転移性」といった疾患の性質を意味する単語に関しては分割して形態素として適用した.

6. まとめと今後の課題

本稿では WWW 上に存在するがんに関する情報を解析するために，3116 語からなるがん専門用語辞書を作成した．そして，blog ページに出現するがん専門用語を抽出し，作成した辞書と照らし合わせた結果約 65%はカバーできた．今後は，がん専門用語辞書に登録するべき用語を明確に決定し，再度調整していかなければならない．そして，WWW 上のがんに関する情報を整理するために，係り受け解析や情報検索の技術に役立てていきたいと考えている.

謝　辞

本研究を行うにあたり御助言を頂いた国立がんセンター若尾文彦医長，情報通信研究機構竹内友木子氏，ならびに関係各位に深謝する．また，本研究は情報通信研究機構運営費交付金（情報通信部門）,平成 17 年度厚生労働省がん研究助成金研究総合研究「がん情報ネットワークを利用した総合的がん対策支援の具体的方法に関する研究 」若尾班等の支援を得て行った.関係各位に深謝する.

文　献

[1] 中川晋一，木村俊也，三角真，島津明，山岡克式，酒井善則，介入的手法によるがん情報取得適正化に関する検討，DEWS 2006 Proceedings

[2] 木村俊也，中川晋一，三角真，島津明，山岡克式，酒井善則，がん情報 Web コミュニティ形成のためのコンテンツ空間の検討－Bayesian classifier を用いたがん情報コンテンツの分類－, DEWS 2006 Proceedings

[3] 国立がんセンター（がんの統計'05）
http://www.ncc.go.jp/jp/statistics/2005/index.html

[4] 厚生労働省
http://www.mhlw.go.jp/

[5] 松本裕治，北内啓，平野善隆，松田寛,"形態素解析システム「茶筅」version 2.3.3 使用説明書"，奈良先端科学技術大学院大学松本研究室 2003 年 8 月

[6] 検索エンジン goo
http://www.goo.ne.jp/