がん用語集合の作成とその特性

中川 晋一†‡* 内山 将夫‡ 三角 真‡ 島津 明† 酒井 善則*

‡情報通信研究機構 〒184-8795 東京都小金井市貫井北町 4-2-1

†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

*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: ‡{snakagaw, mutiyama, misumi}@nict.go.jp †{shimazu}@jaist.ac.jp, *ys@ss.titech.ac.jp

あらまし ウェブ等で提供されるがん情報を理解し，適切に患者や家族に提供する事は非常に重要である．その補助のための文書分析を可能にする基盤であるがん用語辞書を，医師免許を持つ専門家による人手で作成した．がん用語抽出の対象として，権威あるコーパスである国立がんセンターのウェブ文書全体（59 のがん疾患に関する説明を主とする）を対象とした．のべ約 2 万 6 千の用語を収集し，用語候補の集合 Cc（Cancer Words Candidate：語彙数 10199 語）を得た．網羅率は，数種のがん説明用コンテンツを対象として 94.7%から 99.5%であった．一般語やがん医学用語との関係ならびに用語集合内の意味の整合性から用語選択の基準を作成し，その基準に基づき Cc から用語を抽出したところ，93.7%が基準に合致した用語であった．また本がん用語辞書は，従来汎用されてきた医学用シソーラス（MeSH）とは異なる概念である症状，病名，診断，治療，予防などの観点からの用語に分類できた．これらより，本用語辞書によって，ウェブ等の文書から，がんに関する情報を適切に収集できることが期待される．

キーワード 情報検索，文書分類，用語辞書，単語集合の評価

1. はじめに

がんの患者や家族にとって，がんに関する情報(以下，「がん情報」と呼ぶ) を知ることは非常に重要である．その情報源として，専門的で高価な医学書に比べ，ウェブで提供されているがん情報（以下，ウェブがん情報と呼ぶ）は容易に入手可能であること，用いられる用語が平易であることから広く用いられるようになってきた[1][2]．しかし、これらウェブがん情報は，良質なものばかりではない．例えば，「肺がん」を検索語として Google で検索した場合，検索結果の中には，信頼性の高いページもあれば，医学的な根拠がなく高額な民間療法への誘導のページもある．そのため，がん患者がこのように悪質な情報により，被害に合う場合もあることが問題である．これら公開されているがん情報は，根拠のある科学的なものと，有用であると個人が主張するもの，さらに商用誘導まで存在する[3][4].

このようながん情報のなかから，良質ながん情報を選別し，その文章が何を述べているかを推定する必要がある．例えば以下のような文章，

文例１：がんの記録の例「患者は，１年前に左肺上葉切除ならびに肺門部リンパ節郭清を行った．術後１０ヶ月経過時点で CT 上肝右葉への遠隔転移を認めたため，CDDP+エトポシドの化学療法を行ったところ転移巣は消失した」

から，

1. 「CDDP」と「エトポシド」が，がんの薬名である．
2. 「左肺左葉切除」が肺がんの手術の方法である．
3. 「転移」が，がんに特有であること
4. 「転移巣」から転移を起こした進行がんであること．

を理解したい．そのためには，少なくとも，「CDDP」「エトポシド」「左肺左葉切除」「転移巣」「転移」ががんに関する用語(以下，がん用語と呼ぶ．)であることを知るとともに，それぞれの語の意味を知る必要がある．ところが，がん専門用語を網羅した辞書は，言語に関わらず印刷物でも存在しない．本研究でいうがん用語の一部，例えば「リンパ節」や「転移」は，一般用語辞書(例えば ChaSen 用の ipadic ver2.7.0) や医学用シソーラスである MeSH[5]にも含まれているが，これらの語が，がんに関連する用語であるとの説明がないため，意味を理解する事はできない．これら理解をするために，がん用語集合を作成する必要がある．

このため，網羅的に幅広く候補語を収集し，どの語ががん用語として適当かを検討し，選択基準を作成することで，がん用語集合を定義してゆく必要がある．本研究では，まず有資格者の語感を用いて，権威あるコンテンツから網羅的に用語を収集する（がん用語候補集合の収集）．各用語の意味と他の用語との比較から，がん用語の用件を検討し，実際の用例に基づいてがん用語集合の要件を明らかにする（がん用語選択基準の作成）．さらに，本用語集合に含まれる用語の意味内容を分類し，本用語集合の意味内容の特徴を示す．

2. がん用語辞書の必要性

厚生労働省の人口動態統計をもとに作成した死亡率の推移を図 1 に示す．わが国のがん患者数は 100 万人以上であり，家族をあわせると，がんに関する情報検索を行う可能性のある人口は数百万人に上る．検索対象として極めて重要である．

また，カルテには，検査，所見，患者の訴え，主治医の考えが時間経過で記述され，例えば「ベッドから転落して骨折」

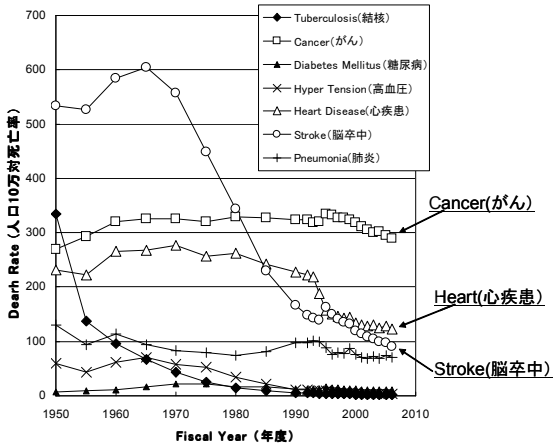

図 1 ：死亡原因の変化（死亡数/人口 10 万人/年）

等の，がん以外の部分と，がんの内容が混在する．例えば，骨折に対する抗炎症剤とがん疼痛のための薬剤が同じであるため，骨折治療のために，がん疼痛の治療のために何の薬剤が使用されたかを混在する文書中から逐次検索しなければならない．単に薬剤名の検索だけでは薬剤の使用目的が不明であるため，一連の記述の中からがん疼痛に対する治療を行っている記述を選択的に提示することができれば，重複投薬による医療ミスの軽減に対して有効である．そのために，例えば出現するセンテンスの中で，がん用語の頻度を計算するなどにより，がんに関する部分を分離する事が可能になる．

専門用語を収集するアルゴリズムは従来研究でも行なわれている[6][7]．しかし，作成された専門用語候補の中から，どの用語を選択するかに関しては，もう一度コーパスの文脈を参照する作業を人手で行う必要がある．そのため，最初から人手で用語を収集することにした．

3. がん用語候補集合（Cc）の作成

以上から本研究では，以下の手順で用語辞書を作成した．概略を図 2 に示す．1. コーパスの選定：信頼性と網羅性の高いがん情報に関するコンテンツを選定（図 2 左①-1，①-2）2. 専門家による用語の収集：専門家の語感を信用し，網羅的に人手でがん用語候補を収集する（C1,C2→Cc）3. Cc を吟味し，用語集合としての整合性と他の用語との関係等を検討して選択基準を作成する（図 2 右②）．

図 2 の①に示すように，2006 年 6 月に国立がんセンターの情報提供を行っているコンテンツのうち，がんそのものに関して説明を行っている 53 疾患（図 2 左：「各種がんの説明のページ」，約 1.5 メガバイト）を，それぞれの疾患の説明別にテキストファイルを作成した．このテキストファイルの中に対して，専門家が，「がんに関連する用語」として認識する用語を幅広く網羅するように名詞句を中心として切り出し，がん用語候補集合 C1 を作成した．

これらコンテンツは，1 ページあたり約 2000 字から 15000 文字，ファイルサイズとして 10K（5000 文字）から 30K バイト（15000 文字）であり，切り出しに要する時間は 10K バイトあたり約 30～45 分を要した．切り出し語数は 1 疾患あたり 150～350 語であった．得られた C1 の異なり語数は 3313 語であった．C1 作成時の語彙の成長曲線（growth curve）を図 3 に示す．さらに同年 10 月に行われた大幅改訂に伴い，疾患数も追加され，53 から 59 となった．そのため，より網羅性を高くすることと関連用語も含めた用語収集を目的とし，疾患別だけではなく全ページ（データ量は合計約 250M テキストとして容量約 15 メガバイト）を対象として，延べ 29500 語を切り出し，用語集合 C2(9451 語)を得た．このようにして得られた用語集合 C1，C2 の和をとり，がん用語候補集合 Cc(Cancer words Candidate, 10199 語)を得た．

4. Cc の特徴と選択基準に関する検討

実際の用語切り出し例の特徴を検討した． Cc に含まれる各用語は，「肺，心臓，気管支」などの解剖学用語，がんの専門知識を表現する用語，「罹患率，死亡率」などの疫学用語，「転移，腫瘤」などの病理学用語，「肺の構造」，「肺がんの統計」などの元コンテンツでの見出し語など様々な語句が含まれる．さらに「がんの発生」，「遺伝子の異常」などの「用語 A＋“の”＋用語 B」や，「他覚的な副作用」，「白血球減少」，「肝機能障害」などの複合語も含まれている．これらは，

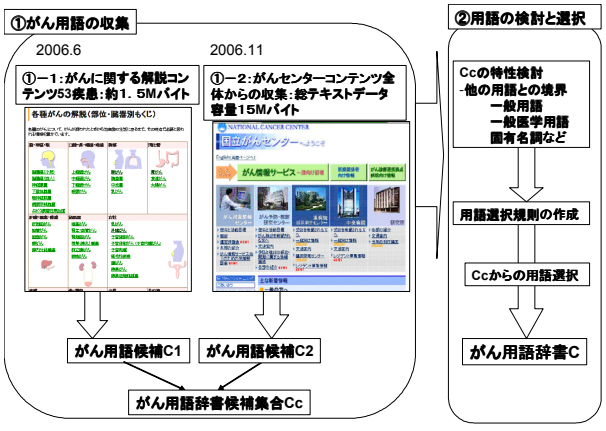

図 2 ：本研究でのがん用語の収集と選択の概要

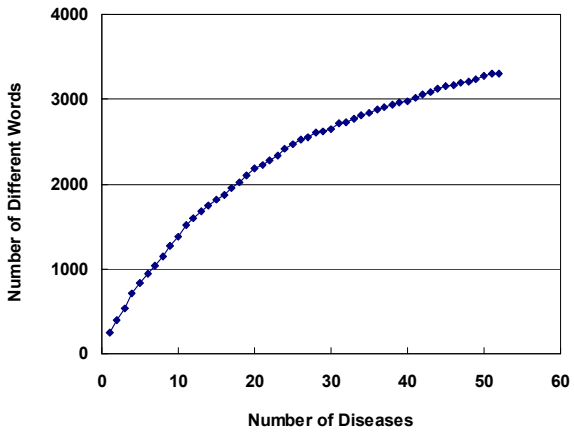

図 3 ：C1 作成時の疾患数別 Growth Curve

がんに関しての中心的な用語であり，「がんの事典を考えたときに，単独の項目を構成することが可能であるような用語」であった．

また，これら用語は網羅性を重視して関連用語も収集されている．そのため，文脈上の用例を再度検討することによって「減少傾向，国際比較，日本人，欧米人」等の通常の一般名詞句や，「空気，酸素」等の医療行為に用いるが，がんと直接の関係のある文脈で使用されていなかった語，「正常の機能，発生」等の医学用語だが，単独で出現した場合，がん用語として扱う事が望ましいとは限らないもの，「肺内」のように場所の概念そのものがはっきりとした概念になりにくいものなど，除外すべき用語も含んでいた．

これらのことから，Cc に含まれるがん用語には，従来の用語辞書では網羅できない，がん特有の用語を含んでいる事，がんに特有な解釈から関連用語とする方が望ましい用語，さらに再検討によって除外すべき用語も存在することが示唆された．これらの差異は語感に頼るしかないが，語感のみではがん用語の選択に揺れが生じる可能性が高い．例えば，がん予防の面から食品名（うなぎ，にんじん）が列挙される場合も多いが，「うなぎ」をがん用語に含めた場合に「魚」もがん用語として含めるかなどの問題が生じる．そこで，それら境界を「２ホップまでの原則」として設定した．これら用語と，一般用語や一般医学用語などとの関係に関しては，後に詳しく検討する．以上により得られた，がん用語選択基準を以下に示す．

がん用語選択基準

1. がんの事典を考えたときに，単独の項目の構成が可能

例：　肺がん肝転移：原発性肺がんの肝臓への転移
　　がんの発生　：がんが生じること

2. がん関連用語も採用する．

根拠が十分　：疫学用語でリスク評価指標
β－カロテン：がん予防の可能性のある栄養素

3. ２ホップまでの原則：がんの関連用語の関連用語までは採用するが，それより関連が薄いものは採用しない

１ホップ目：β－カロテン：がん予防栄養素：採用
２ホップ目：にんじん：β－カロテン食物：採用
３ホップ目：野菜：健康増進一般に関係する：不採用

5. がん用語集合の特性

得られた Cc（10,199 語）を対象として，用語収集の一貫性とがん用語選択基準の妥当性を，以下により検証した．

5.1. Cc 収集の Consistency

作成した用語集合 Cc の網羅性（カバー率）を検討するため，まず，ALL（急性リンパ急性白血病），腎細胞がん，膵がん，卵巣がん,肺がん，肝臓がん,グリオーマ,胃がん,大腸がん,乳がんを対象として，原文書であるがんセンターのコンテンツから，約 1 年後 Cc を収集した人と同一人物が，3. と同様に人手でがん用語を抽出した．次にその用語集合と Cc を比較し，Cc がどの程度抽出された用語を網羅しているかを確かめた．ここで，この用語集合と Cc とは，同一人物により抽出されたものではあるが，1 年の間隔が空いているため，がん用語の網羅的収集ということを目的として，独立に採集されたものとみなせると考えている．そのため，もし，Cc がこの用語集合を広く網羅しているとすれば，Cc は，がん用語を広く網羅すると考えられる．つまり，Cc は，医師の有資格者の語感による，がん用語の候補を広く網羅すると考えて良い．結果を表２に示す．これら原文は 2007 年現在（Cc 作成から約 1 年後）のもので，Cc を作成した時点と若干の変更が加えられている．

表２の①は，それぞれの疾患別に抽出された用語集合の語数である．②は，それぞれの疾患別の用語集合で，Cc に含まれている用語の数である．これらより，③に示したように，見かけのカバー率が算出される．しかし，今回新たに人手で切り出した結果も網羅性を高くすることを意図していたため，すべての用語が必要かどうかは不明である．つまり，今回抽出した用語候補にも，用語として不適切なものがある可能性が高い．そこで今回抽出された用語の中で Cc に含まれていない用語を選別し（④），それら用語の中で 4 節の基準に相当するかどうかによって真に必要な用語数⑤と，選択されたが不必要であった用語数⑥を求めた．これより，それぞれの文書から抽出されるべきだった用語数⑦を求め，②を分子として求めた真のカバー率が⑧である．以上により，これより Cc のカバー率⑧は 94.7%から 99.5%と算出された．

表 2：Cc（10,199 語）のカバー率

病名	①専門家が手で抽出した語数	②:①のうちCcに含まれた語数	③:みかけのカバー率:②/①	④:①の中でCcにない語数	⑤:④中で採用すべき語数	⑥:④中で不要だった語数	⑦:①のうち採用すべきだった語数:②+⑤	⑧:真のカバー率:②/⑦
ALL	368	291	0.791	77	12	65	303	0.960
relancellcancer	208	173	0.832	35	1	34	174	0.994
PancreasCancer	228	187	0.820	41	7	34	194	0.964
OvaryanCancer	280	224	0.800	56	5	51	229	0.978
LungCancer	475	396	0.834	79	22	57	418	0.947
LiverCancer	296	256	0.865	40	10	30	266	0.962
Glioma	236	204	0.864	32	5	27	209	0.976
GastricCancer	541	437	0.808	104	19	85	456	0.958
ColonCancer	517	432	0.836	85	8	77	440	0.982
BreastCancer	346	302	0.873	44	10	34	312	0.968

ALL：急性リンパ球性白血病，renalcell cancer：腎細胞がん，PancerasCancer：すい臓がん，Ovaryan Cancer：卵巣がん，Lung Cancer：肺がん，Liver Cancer：肝がん，Glioma：グリオーマ（神経膠腫），Colon Cancer：大腸がん，Breast Cancer：乳がん

5.2. 用語集合 C の選択と他の用語集合との境界

用語候補の集合 Cc から妥当ながん用語集合 C を得るために，Cc 10,199 語を ASCII コード順にソートし，医師免許を保有する有資格者（Cc 作成者と同一人物だが，Cc を作成してから約 1 年経過後，Cc 作成時と独立の判断ができると仮定した）と自然言語処理の研究者 1 名が，がん用語選択基準に基づいて概念範囲と語彙整合性を整理しつつ，1 語 1 語を音読し判断した．各用語を図 4 に Cc に出現した肺がんを例とした選択基準による用語選択の概念を示す．肺がんを基点とする場合、扁平上皮がん、腺がんなど肺がんそのものの分類による名詞句を 0 ホップ目とする．この場合，治療に使用される「ブレオマイシン」，手術法の「左肺切除術」，転移の病態を示す「肺がん肝転移」，原因とされる「タバコ」，「アスベスト」を 1 ホップ目とする．「ブレオマイシン」は「抗生物質」の一種であり，「抗生物質」は医学一般では「肺炎」の治療に使われる．このように，肺がん-ブレオマイシン-抗生物質-肺炎という連関を想起しつつ，2 ホップ目までの想起語をがん用語として採用，3 ホップ目である「肺炎」，「気胸」，「肝膿瘍」は Cc から削除するという手順で用語を選択した．この他，医学用語は伝統的に発見者に敬意を表して固有名詞に単語が付加される合成語が多い．例えば，がん研究・治療の中核的機関である合衆国の「M.D. Anderson Cancer Center」など個人の名前を冠した公的機関も多い．これらの固有名詞を含む用語の中で，白血病の分類基準である「WHO 分類」は，WHO（世界保健機関）が決めた分類の名称であり，がん用語だが，「WHO」（世界保健機関）は固有名詞である．以上により，Cc から 640 語を除去し，合計 9559 語を用語集合 C として選択し，用語候補 Cc の 93.7%を用語として採用した．

5.3. がん用語の特徴と今後の課題

用語集合 C から無作為に用語 100 語ずつを選択し，一語一語上から順番に「低血圧」なら「症状」という分類概念を作成し，以降「耳鳴り」「イライラ感」などの用語が出現するたびに「症状」に分類する．その後，「胃がん」と出現すれば，「病名」という分類概念を作成，「胃潰瘍」が出現した

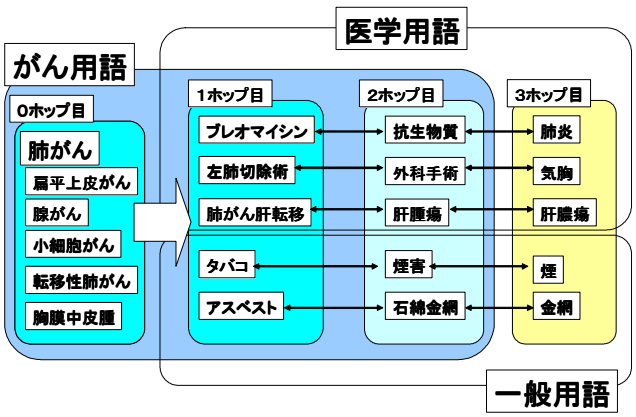

図 4：がん用語，一般用語，一般医学用語の関係の概念図

図 5：C から抽出されたがん用語の分類概念

場合は「病名」を「がんの病名」と「関連疾患病名」に細分化する．以上のようにして，分類概念を作成しつつグループ化した．数回繰り返し，「症状」や「病名」の 5 つの大分類と「がんの病名」など 11 の項目に分類されることを見出した．結果を図 5 に示す．分類は，医学の一般常識に合致した．有用と考えられる．今後これらを用いた高度処理について検討する予定である．

6. まとめ

がん情報処理を補助することを目的とし，基盤であるがん用語辞書を，医師免許を持つ専門家による人手で作成した．権威あるコーパスとして選択した国立がんセンターから，人手で抽出を行い，10199 語の用語候補集合 Cc を得た．がん用語の基準として，がんに直接関係する用語，がんに関連する用語，がん関連用語の関連用語までを採用する，という「2 ホップの原則」選択基準を作成した．Cc のがん用語のカバー率は 96.5%から 99.5%であった．この基準によって Cc の 93.7%(9559 語)の用語が採用できた．本用語集合と他の用語集合の関係も検討し有用であると考えられた．

謝　辞

本研究は NICT 運営費交付金（新世代ネットワーク研究センター），平成 19 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] 山室真知子，医学情報の患者へのバリアフリー，情報の科学と技術，50(3)，pp138-142，2000

[2] 野口迪子，医学書を探す：基本図書を主として，情報の科学と技術，50(11)，pp542-552，2000

[3] Wendy A. Weiger（原著），坪野 吉孝（翻訳），がんの代替療法―有効性と安全性がわかる本，法研，2004

[4] Humphrey SM, Miller NE. Knowledge-based indexing of the medical literature: the Indexing Aid Project. J Am Soc Inf Sci 1987;38(3):184-96

[5] National Library of Medicine, Medical Subject Headings(MeSH) fact sheet, 2006

[6] H. Nakagawa: "Automatic Term Recognition based on Statistics of Compound Nouns", Terminology, Vol.6, No.2, pp.195 - 210, 2000

[7] 佐藤理史，佐々木靖広．ウェブを利用した関連用語の自動収集．情報処理学会研究報告，2003-NL-153，pp.57?64，2003.