ウェブがん情報空間推定のための単語集合に関する検討

中川 晋一[†‡*] 木村 俊也[†] 三角 真[‡] 島津 明[†] 山岡 克式[*] 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, misumi}@nict.go.jp †{s-kimura,shimazu}@jaist.ac.jp,
*{nakagawa, yamaoka, ys}@net.ss.titech.ac.jp

あらまし Web で提供されているがん情報は、専門用語や一般用語、英語、日本語長文節語などからなること、疾患の特殊性から一般用語辞書や、医学専門用語辞書では形態素解析が困難であった．これら問題を解決しコンテンツの質的評価を行なうため、現在知られ一般に認められ参照される国立がんセンターWeb から、59 疾患（胃がん、肺がんなど）を説明している部分から約 3650 語(Dic-C1)を切り出した．さらに昨年 10 月に改訂された同ＵＲＬ（総量 250M バイト、HTML テキスト 15M バイト）から 15 文字程度までの長文節語（例えば「右開胸開腹胸部食道全摘胸骨後頚部食道胃管吻合３領域郭清」という手術法 27 文字）までをも含む用語辞書を 9650 語を切り出した（Dic-C2). これらと一般用語辞書(i-dic:約 7 万語)、提供される一般医学用語集約 6 万語(Dic-M)の 4 種類の単語について、検索エンジンで得られた 59 種がん 44777 URL における出現頻度、重複度等に関して検討し、Dic-C1,C2 の特徴について検討した結果を報告する．

キーワード ウェブマイニング，情報検索，文書分類, 用語辞書，単語集合の評価

1. はじめに

がん患者や家族にとって，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[1]. インターネットによるがんに関する情報発信は活発化し、情報の量が増加してきているが，中川・木村[2][3][4][5]は，各種がんについて URL を外的基準により再評価する必要を報告した．同時に Web 上に存在するがん情報は千差万別であり、患者にとって必ずしも有用な情報を提供しているものばかりではないことも示唆した．

1.1. がん情報分類のための専門語集合の必要性

一般に Web 等で提供されているテキストデータを機械分類を用いて評価しようとする場合、TFIDF、T/R 比等の評価基準による外的尺度が検討されてきた．木村・中川らの検討[3]は、一般用語辞書を用いたベイズ分類器による実験により、がん情報を提供している URL を分類、分類アルゴリズムは実用レベルにあることが示された．しかし、分類されたグループにおける一般用語集合（i-dic：約 7 万語）の中での単語は「私、先生」等の人称語等の疾患概念に関係のないものも多かった(Table 1). 最適分類を与えるだけでは不十分であると思われた．これは、それぞれの URL の発信者の背景に保持する言語空間が異なっていること、発信者が想定する対象が異なっているなどが考えられる．概略を Fig.1 に示す．例えば，Wg: 一般用語、Wm: 医学専門用語に加えて、がんの特殊な言葉集合 Wc が存在することが考えられる．本研究では Wc という言葉集合を作成、Wg, Wm 言葉集合と特性を比較検討することにより、がん情報発信を行なっている URL の言語空間の特性評価に関しての有効性、妥当性を検討することとした．

Table 1: Characteristic words CII=2 and CII={1,3,4} groups

Word	Difference	Word	Difference
私	7216	研究	−4987
入院	3917	相談	−4558
病院	3905	漢方	−3888
検査	3240	シート	−3066
自分	3214	情報	−2086
先生	2336	一覧	−2069
海外	1875	抗がん剤	−2062
手術	1871	内容	−2034
これ	1816	必須	−1739
人	1805	薬局	−1599

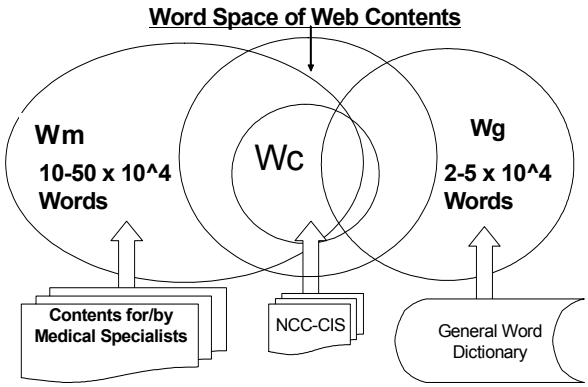

Fig.1 Assumed Relations and Scales of Each Word Set for Cancer Information Providing

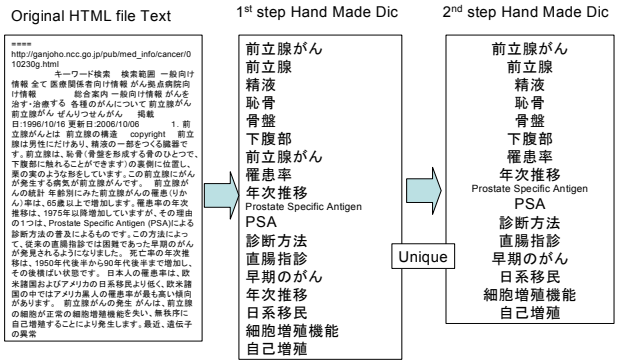

Fig.2: Process of Making the Hand Made Dictionary from URL Contents.

1.2 がん用語の特殊性

がんは、高血圧や糖尿病のように治療法の確立している疾患群とは異なり、医師にとっても特殊な用語が存在する．特に、治療方針を説明し同意を得る「インフォームドコンセント」という過程が不可欠であり、その説明のために医師も患者に対して特殊な言葉遣いをすることが多い．例えば「転移性肺がん」という文節を従来の用語辞書で次のように切り出しては意味がない．

転移性肺がん
- 転移
- 性
- 肺
- がん

両側性肺門部リンパ節腫脹
- 両側
- 性
- 肺
- 門
- 部
- リンパ節
- 腫脹

「転移性肺がん」「両側性肺門部リンパ節腫脹」それぞれ、一文節一語として頻度を算出する必要がある．そのため､未知語を採集する URL が決まっても形態素解析を行なう場合、通常の一般用語辞書を用いて既知語の「転移」「性」「肺」「がん」に分類されると結果を誤る．専門用語を抽出するアルゴリズムも提案されており[7]、実装し抽出を試み約３万語を得たが、誤抽出も相当数（約２割程度）あり、専門的知識を有する有資格者（医師）が直接切り出すほうが効率的であると判断した．長文節の単語からなる辞書を Fig.2 に示すように直接作成する事とした．

2. 専門用語辞書の作成

2.1 Wcc(がん専門用語辞書)の作成経過

国立がんセンターの Web ページ[9]の文章を元に用語辞書を作成した．がん専門用語は，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．作成時、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（W_{c1}-W_{c54}）を切りだした．それぞれ和をとり異なり語集合を求めワードセット C1-Dic とした。

Table 2.: 59Cancers of this study.

胃がん	菌状息肉症	中皮腫
肺がん	形質細胞性腫瘍	聴神経鞘腫
大腸がん	原発不明がん	軟部肉腫
肝臓がん	喉頭がん	尿管がん
白血病	骨髄異形成症候群	脳腫瘍
乳がん	子宮体部がん	皮膚がん
子宮がん	子宮肉腫	非ホジキンリンパ腫
ぶどう膜悪性黒色腫	子宮頸部がん	慢性リンパ性白血病
ホジキン病	上咽頭がん	慢性骨髄性白血病
悪性リンパ腫	食道がん	慢性骨髄増殖性疾患
悪性黒色腫	神経膠腫	網膜芽細胞腫
咽頭がん	腎細胞がん	卵巣がん
陰茎がん	腎盂がん	卵巣胚細胞腫瘍
下咽頭がん	成人T細胞白血病リンパ腫	睾丸腫瘍
下垂体腺腫	精巣腫瘍	絨毛性疾患
外陰がん	前立腺がん	膀胱がん
肝細胞がん	多発性骨髄腫	膣がん
急性リンパ性白血病	胆管がん	膵がん
急性骨髄性白血病	胆嚢がん	膵内分泌腫瘍
胸腺腫	中咽頭がん	合計 44,910URLs

2.2 C1-Dic 作成経過

疾患別に用語集合 W_{ci} を加えたときの C1-Dic 内に存在する用語数の変化を Fig.3 に示した．縦軸は C1-Dic の語数である．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 3313 語となった．Fig.3 のがん毎の微分値のプロットを Fig.4 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.5 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．これらから各疾患を解説するのに用いられる用語は多くが重複していることが示唆された．以上がん用語辞書”C1-Dic”（3313 語）を固定した．

さらに 2006 年 10 月に国立がんセンターホームページが大幅に改訂された．ダウンロードしたデータ量は合計約 250MB であった．この中から、HTML ファイルのうち、テキストデータのみを抽出し、合計 15M バイトを得た．ここから疾患数も追加され、53 から 59 となったため、本研究で検討するがんの種類も Table 3 に示す 59 種類(Table 2)とした．再度同様の手法で”C2-Dic”(9451 語)を作成した．

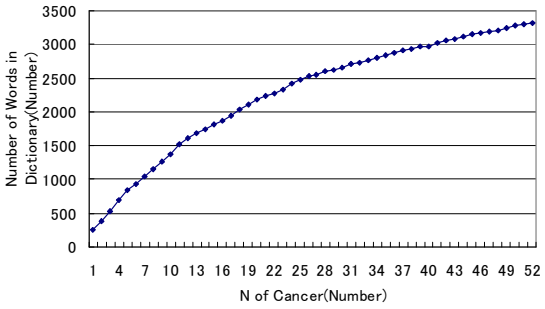

Fig.3: Number of words in C1 Dictionary

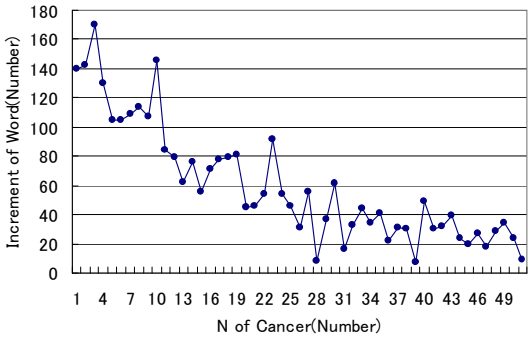

Fig.4 Number of addition by each cancer

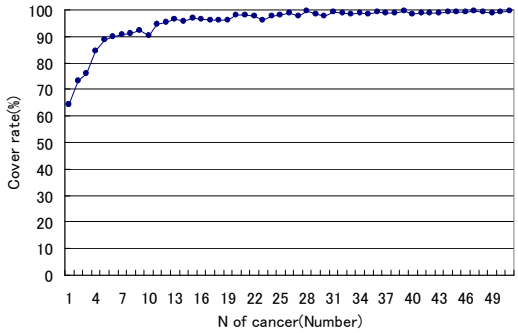

Fig.5: Cover rate of cancer dictionary by Number of Cancers

3. がん用語辞書 C-Dic の評価

3.1 比較のための単語集合の設定と構成の概要

C1, C2 に含まれる語彙の特性を検討するため、一般用語辞書(i-dic: 75498 語)、インタネット上で収集した複数の医学用語辞書(M-Dic: 59533 語)を比較用の語集合とした．Table 4 および Fig.6 に語集合 G,M,C1,C2 におけるそれぞれの単語長の分布を示す．最長単語長は

Table 3. Basic Statistics of G, M, C1 and C2 Word Set.

	Category of Dictionary			
	G	M	C1	C2
N of words	75498	59533	3315	9451
Averaged Length of Word	2.90	10.14	4.90	5.86
S.D. of Length of Words	1.26	7.15	2.17	4.43
Min Length	1	1	2	1
Max. Length	13	79	35	89

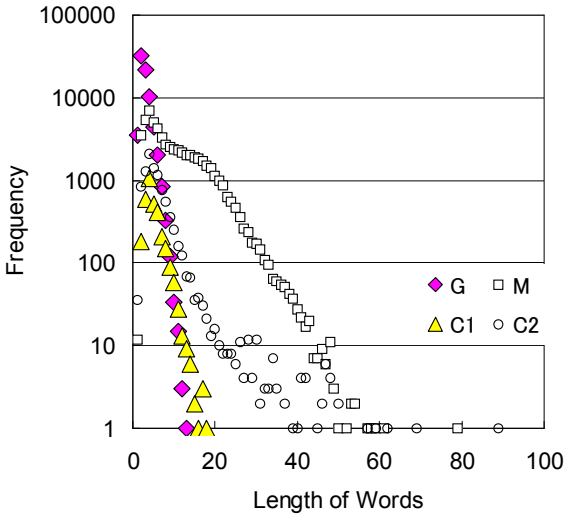

Fig.6: Comparison of G, M, C1 and C2 Word sets (Word Length – Frequencies).

G: 13, M: 79, C1: 35, C2: 89、平均単語長 G: 2.9, M: 10.4, C1: 4.9, C2 5.9 であり、医学系 3 種類の M,C1,C2 が一般用語集合 G に比べて長い．

Fig. 10 より、G に比べ、M が 10 文字以上の単語が多い事、C1 より C2 が長い語を含んでいる事がわかる．M は複数文節からなる英単語を含んでいる事、C2 は C1 に比べ、より精密ながんに関する情報を提供する目的のために改変されたコンテンツからなっていた事から、C2 の語数が多くより専門性の高い M に近い語彙を含んでいることが推測された．以下、これら単語集合それぞれの重複語彙数を Table 4 にまとめた．

3.2 対象とするがんＵＲＬ集合と解析用データの固定

Table 2 に挙げた 59 疾患それぞれを検索語として、検索エンジン(Google!)を用いて URL リスト(それぞれの疾患において約 800－900 個程度)を取得、合計 44477 個の Valid な URL リストを得た．本 URL リストをもとに、wget を用いて、同数の URL ツリーを全量ダウンロードし、HTML ファイルとして合計 1.18 ギガバイトを得た．本データを対象として、G, M, C1, C2 それぞれの重複語集合の出現頻度と出現 URL 数をカウントした．結果を Table 5 に示す．

3.3 各用語集合の特性に関する検討

3.3.1 出現頻度

各語に関して 1 回以上出現したものを”Appear”とし、重複語数との比を Appearance Rate とした．結果を Table 4 に示した．本 URL 集合において、G に比べて M,C1,C2 の Appearance Rate（出現比率）が高いこと、中でも G の 0.58 に対して C1 の 0.74 は語彙数が G に比べて 20 分の 1 程度であるにもかかわらず高い事から本 URL データでの特異性が高かった．また、C1∩M、C2∩M が、それぞれ語彙数 801 と 1974 に対して C1∩C2∩M が 686 と C1∩M の率が高かった．また、C1∩

Table 4: Result of Survey of 59 Cancer-44477 URLs for Each Word Sets (G, M, C1, C2 and Duplicate Data Sets)

Category of Sets	N of Words	N of Appearance	Appearance Rate
C1	3315	2470	0.745
C2	9451	5826	0.616
M	59533	40381	0.678
G	75498	43942	0.582
C1∩C2	2494	1948	0.781
C1∩M	801	762	0.951
C1∩G	103	101	0.981
C2∩M	1974	1870	0.947
C2∩G	1046	1021	0.976
M∩G	2827	2588	0.915
C1∩C2∩M	686	657	0.958
C1∩C2∩G	66	64	0.970
C1∩M∩G	69	69	1.000
C2∩M∩G	629	626	0.995
C1∩C2∩M∩G	48	48	1.000

C2∩M と C2∩M∩G の語彙数が両方とも 650 前後であり、C1∩C2∩G と C1∩M∩G がほぼ同数の 66－69、出現数もほぼ同程度であったことから C1∩M が特異的に出現していることも推定された.

3.3.2 C1, G を構成する語に関する検討

前節での検討により、C1 が語数の少ないこと、特異度が高いことからがん情報コンテンツの内容把握や評価を与える目的において有用であると思われた. そこで、C1 を構成する語彙の特徴を一般用語集合 G と比較した. C1 と G の単語長の 44,477URL にそれぞれの特徴の概要を Fig6 に示す. C1 は G に比べ、5 文字から 10 文字において、出現回数が 1000 をこえるものがある. C1 では, この単語長では、骨髄性白血病, 悪性リンパ腫,などが出現しているのに対して、G ではアレルギー, クリニック, ランキング,アルコール, カテゴリーなどのカタカナ語であり、両者の語彙集合の質は異なることがわかる. さらに 10 文字以上の語において G ではコミュニケーション, リハビリテーション, インフォメーションなど全てカタカナ語であったのに対して、C1 では非ホジキンリンパ腫, 急性リンパ性白血病, セカンドオピニオンといった、がん情報に関連した語であった. 上記の G に含まれるカタカナ用語と C1 に含まれる語は意味内容と医学的知識への適合性が異なっており、がん用語辞書 C1 を用いた方がより理解しやすいと考えられた. 以上のことから、語集合 C1 によりがん情報 URL は一般用語辞書を用いた場合に比べて的確に内容を把握できることが示された.

まとめ

がん情報提供状態の質的評価を目的としたがん専門用語集合について検討した. がん用語集合 C1,C2 は一般に認められ広く用いられている国立がんセンター(NCC-CIS)コンテンツから手で切り出して作成し (C1、C2)、C1:3316 語、C2 9475 語を得た. 本言葉集合は一般用語辞書の約 7 万、医学用語辞書 M の約 6 万に比べて小さかった. C1,C2 の特性を検討するため、各種が

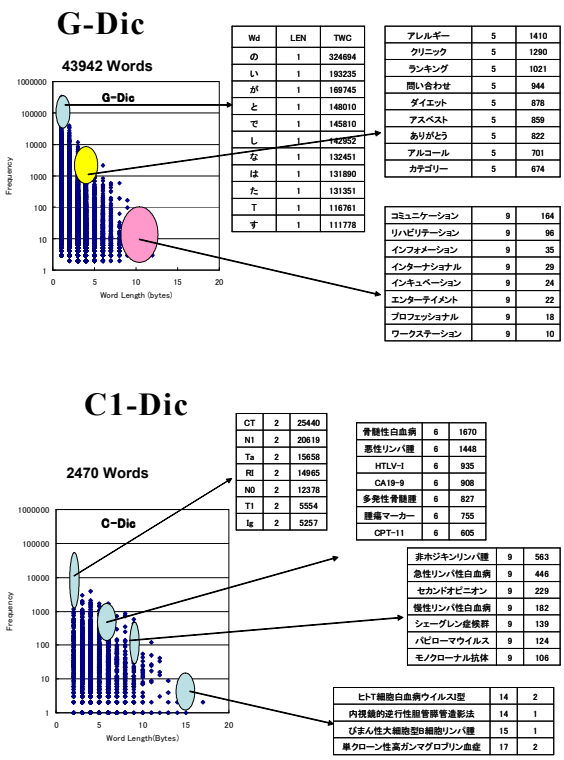

Fig.6: Detailed Comparison between G and C1 by Frequency of 59 Cancer (44,477 URLs)

んを検索語として既存の検索エンジンで得られる URL リスト(44777 個)における出現頻度を測定した. その結果、C1 は一般用語辞書 G の 58%に比べて約 75%と高く、本集合がん情報コンテンツの内容に特異性が高い事が示された.

謝　辞

本研究は情報通信研究機構運営費交付金(新世代ネットワーク研究センター), 平成 18 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] NHK SPECIAL HOME PAGE, http://www.nhk.or.jp/special/libraly/06/10001/l0107.html
[2] 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, DEWS2006 Proceedings, 1b-i10, 2006
[3] 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－, DEWS2006 Proceedings, 1b-i9, 2006
[4] 木村俊也, 中川晋一, 三角真, 山岡克式, 酒井善則, 島津明, Web 上のがん情報取得のためのがん用語辞書の作成, NLP2006 Proceedings, 2006
[5] 中川 晋一, 木村 俊也, 三角 真, 島津 明, 山岡克式, 酒井 善則, 患者のためのがん情報 URL リスト適正化に関する検討, DBSJ-Letters Vol.5 No.1, pp21-24, 2006
[6] Hiroshi Nakagawa: "Automatic Term Recognition based on Statistics of Compound Nouns", Terminology, Vol.6, No.2, pp.195 - 210, 2000
[7] 国立がんセンター http://www.ncc.go.jp