一般向け医学文章の平易化を目的とした医学専門用語の自動獲得

山田恵美子*　　荒牧英治**　　大江和彦*
*東京大学大学院医学系研究科　　**東京大学医学部附属病院
emiko-tky@umin.ac.jp
{aramaki, kohe}@hcc.h.u-tokyo.ac.jp

1. はじめに

近年医療に対する国民の関心は高まっており、一般書籍やインターネット等のメディアにおいて、医学知識を持たない一般人を読者対象とした医学文章が存在する。ところがこれらの医学文章を執筆するのは医学知識を持つ人間であり、作成された文章は必ずしも一般人にとって読みやすいものとは言えない。そこで我々はこのような医学文章を一般人が理解するための支援を目標として研究を行っている。読解支援をするためには(1)まず文章の中で難解な部分を特定し、(2)そのうえで各部分に対して解説文へのリンクを付与するなどの支援を行うことが必要である。

本研究では(1)難解な部分の特定に焦点を当てる. 文章を難解にする要因としては次の 3 種類が考えられる[5]： (a)語彙的要因、(b)形態・構文要因、(c)談話的要因。

特に医学医療分野では、文章を難解にする原因として(a)が占める割合が 48.2%と大きい[9]。そこで、本稿では語彙的要因のうち医学医療分野で特に問題となる難解な医学専門用語に焦点を絞り、その獲得をこころみる.

2. 難解な専門用語

用語をその専門度合と難解度合という二つの尺度で捉えると，図１のように分けられる.

難解さ

まず、用語を難解度合で分類することができる。例えば「クレアチニン」は一般人には難解であるが「肺がん」は平易であろう。そこで、これら区別することが必要となる。

専門性

次に用語は専門用語と一般用語に分けることができる。玉村[8]は「専門用語」の条件として、

(1) 使用分野、使用者の職業などが限定されている語
(2) 当該分野ではほぼ一義的に理解されるようになっている語

を挙げている。

医学専門用語として一般に認識されている語を集めたリソースとして医学辞書や各種用語集・オントロジーが挙げられる。しかし、これらが含む用語は必ずしも上記の専門用語の条件(1)を満たすわけではない。例えば医学書院医学大辞典には「HTML」「医師数」といった一般用語が掲載されている。このように医学用語と一般用語の集合はオーバーラップすると考えられる（図１）。このような医学用語（専門用語）ではない語は、難解であっても本研究で扱う対象外であり、区別が必要である。

このように難解さと専門性という２つの視点から分類が必要であり, 次章の方法でそれを試みる。

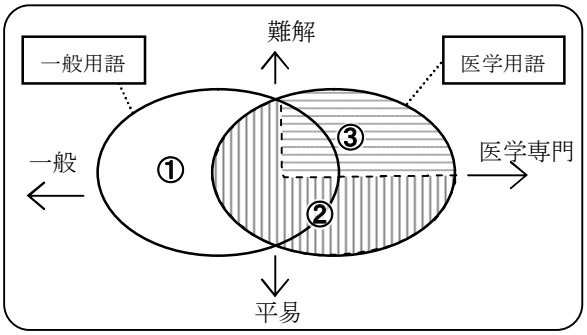

図 1. 用語の分類
①一般用語、②医学専門用語、③難解な医学専門用語

3. 検索 Hit 数と辞書による分類

前章で述べたとおり、問題は②医学専門用語と一般用語のオーバーラップ部分と③医学専門用語から難解かつ専門性のある部分を分類することである。我々は②を医学辞書と一般辞書の両方

に収載されている語（以下、BOTH）、③を医学辞書だけに収載されている語（以下 MED）とみなした。これらの関係を表 1 に示す。表中の数字は図 1 で示した数字に対応する。

表 1. 用語の分類

		医学辞書	
		収載	非収載
一般辞書	収載	②非医学用語（BOTH）	①非医学用語
	非収載	③医学用語（MED）	不明

次に問題となるのは、これら（BOTH+MED）から、難解であり専門性のある語だけを分類することである。我々は、用語の専門度合と難解度合は共に語の出現頻度と相関があると仮定し、検索 Hit 数を利用して分類を試みた。検索エンジンとして Google を利用した。

医学辞書として医学書院医学大辞典[4]を、一般用語集として日本語語彙大系[3]を利用した。

4. 結果と考察

提案手法の妥当性をしらべるため、人手で分類を試みた。

4．1　実験データの構築

表 1 の BOTH、MED それぞれにおいて Hit 数が 10^8、10^7、10^6、10^5 から直近 10 語ずつ、計 80 語について、(1)医学専門用語か否か（専門性）、(2)一般に馴染みのある語であるか否か（難解さ）を人手で三段階評価した。評価者には表 2 のような指示を出し、語と医学辞書でのその語の定義文を見て評価を行ってもらった。

評価は○△×の三択で行ったが、医学専門用語か否かについては△を○に、一般に馴染みのある語か否かについては△を×に変えて集計した。

4．2　Hit 数 との比較

今回獲得したい語は医学専門用語かつ難解な語であり、これを満たす語の数を図 2 に示す。BOTH では Hit 数が小さくなるほどに専門用語の割合が増えており、10^5Hit では 70%近い割合で分類が可能である。この結果により一般辞書と医学辞書の両方に含まれている語から医学専門用語かつ難解な語をある程度抽出できるといえる。

一方、MED ではその傾向が薄い。Hit 数は以下のような問題点を抱えている。

(ア) 多義語の問題：「ＰＫ」のように同じ語が異なる意味を持つ場合に上乗せされる.
(イ) 複合語の問題：その語を含む、より長い文字列の Hit 数が上乗せされる
(ウ) 他言語の問題：（中国語や英語等）のページの Hit 数が上乗せされる

MED（医学辞書にのみ掲載されている語）にはこれらの影響を受けやすいものが多いようである。今回人手で評価した上位 10 語は「CAD」「pK」「ILS」「SP1」「CG」「SOAP」「bp」「ウェブ」「MeSH」である。略語は複数の意味を持つことが多く、アルファベットのみからなる 9 語は(ア)(ウ)の影響を受ける。図 2 を見ると 10^8Hit にも難解な医学専門用語が多く出現しているが、本来これらの語の Hit 数はもっと小さな値であると考えられる。これらの影響を消すため Hit 数を調整する必要がある。(イ)(ウ)は検索結果の周辺の語や言語を調べることで解決可能である。（ア)は語義曖昧性（ＷＳＤ）の問題であり、今後の課題としたい。

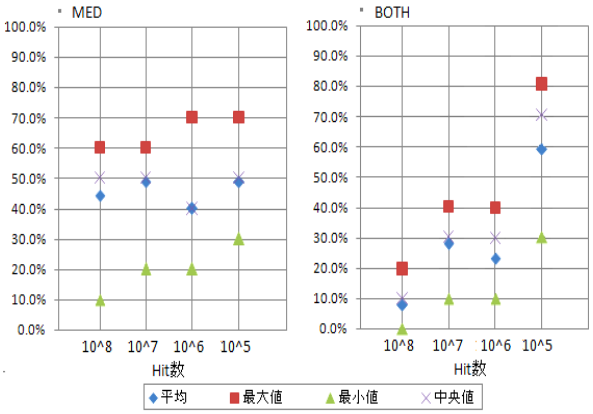

図 2. 難解な医学専門用語の割合

4．3　人間同士の比較

評価者間の一致率は表 3 のとおりであった。同程度の医学知識を持つと考えられる医師間での一致率は他群での一致率よりも低くなっている。これは医学が細分化された分野であり同じ医師でも個々の専門性によって持つ知識や対面する患者の性質が異なるためと考えられる。よって、本研究のような語の分類を人手で行う際には医師よりも一般人の方が統一的な結果が期待できる。

今後の課題

・コンテキストへの依存

本研究では「難解な医学専門用語」を判定することを試みた，しかし，実際の文章中では「難解な専門用語」全てが対象になるわけではなく、コンテキストに依存して決まる場合がある。例えばその文章が「クレアチニン」について説明しているものであればその中で用いられている「クレアチニン」についての読解支援は必要ない。

・医学辞書への依存

医学辞書にのっていない医療用語（例．「免疫システム」、新病、新薬）は対象外となるので他の方法で獲得しなければならない。

5. 関連研究

難解なテキストを平易化する試みがすでに行われている。乾[5]は聾者向けの平易化のため構文的特徴と可読性の関係を調査した。可読性の判断は聾学校教諭によるものであった。聾学校教諭は聾者に知識を与える立場にあり彼らの持つ知識や何を難解と感じるかをよく理解していると考えられるが、医療分野でこれに相当する立場の人間はいない。

新森ら[7]は特許請求項を一般人向けに平易化するための言い換えを試みた。一般的な分かりやすい文章の基本要件に合うように言い換えを行っている。特許請求項はある程度統制のとれた文章であり本稿で対象とする医学文章とは性質が異なると考えられる。

専門用語を収集する方法として特徴語抽出が挙げられる。一般領域のコーパスと対象領域のコーパスでの出現頻度の偏りを特徴度合を示す指標として利用する方法[6]、対象領域のコーパスにおける前後の語の分布を利用して指標を算出する方法[1][2]が提案されている。特徴語抽出は辞書がカバーしていない語を獲得するのに有用だが、対象領域のコーパスを用意する必要がある。日本語の医学領域のコーパスとして整備されたものは存在しない。また、医学医療の話題は専門外の人間でも日常的に扱うものであるため医学領域と一般領域の境界は曖昧であり、コーパスのデザインを困難とする。

6. おわりに

本稿ではHit数と辞書を用いて医学分野における専門かつ難解な用語の分類を試みた。
医学辞書と一般辞書の両方に収載されている語ではある程度の分類が可能であったが、医学辞書のみに収載されている語の分類精度は十分でなく、今後の課題としたい。

謝辞

評価にご協力いただいた鍛治さん、福島さんに心より感謝いたします。

参考文献

[1] Hiroshi Nakagawa, Tatsunori Mori. A Simple but Powerful Automatic Term Extraction Method. Computerm2: 2nd International Workshop on Computational Terminology, COLING-2002 WORKSHOP, pp.29-35,Taipei, August 31,2002.

[2] Katerina T. Frantzi, Sophia Ananiadou. Extracting Nested Collocations. In COLING'96, pp. 41 – 46, 1996.

[3] 池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦．日本語語彙大系．岩波書店, 1997.

[4] 伊藤正男，井村裕夫，高久史麿.医学大辞典. 医学書院, 2003.

[5] 乾健太郎．コミュニケーション支援のための言い換え．言語処理学会第 7 回年次大会併設ワークショップ, pp. 73-78, 2001.

[6] 内山将夫，中條清美，山本英子，井佐原均（2004）．英語教育のための分野特徴単語の選定尺度の比較．『自然言語処理』11⑶：165-197.

[7] 新森昭宏、齋藤豪、奥村学．特許請求項の可読性向上のための自動言い換えのついての考察．言語処理学会第 7 回年次大会併設ワークショップ, pp. 65-70, 2001.

[8] 玉村文郎，『専門用語の性格』，専門用語研究, No. 3, 1-6 (1991)

[9] 山田恵美子，荒牧英治，大江和彦．一般向け医学文章における難解語特定に関する研究．第 27 回医療情報学連合大会.

表 2. 評価者への指示内容

各語について,
1)医学医療用語であるかどうか
2)一般語(医学知識のない人にとっても耳にすることがあり、ある程度理解できる語)であるかどうかを判定し,以下の○△×のいずれかを記入してください.
○:あてはまる　　△:どちらとも言えない　　×:当てはまらない
判定にあたっては,必要ならば語の意味を参考に行ってください.
(例)「プリン」は食べ物の名前として一般語だが、語の意味では物質名で,一般的でないので×)

【声道】 声門より口腔開口端(口唇部)までの口腔,咽頭,鼻腔などの管腔を,音声学的にこう呼ぶ。喉頭の声門部で喉頭原音(声門音源glottal source,声門波glottal wave)が生じ,声道の共鳴によって音色の変化を受けて音声となる。また,フォルマントは声道特性を基に決定される。

【プリン】 C5H4N4,分子量120.11。ピリミジン環とイミダゾール環との縮合環からなる複素環式化合物。本物質は水などによく溶け,その水溶液は中性を示す。天然には遊離では存在しない。

⋮

	GⅠ(医師)		GⅡ(医療関係者)				GⅢ(一般)		
	A	B	C	D	E	F	G	H	I
A		▲0.73 (6.55)	▲0.66 (16.33)	▲0.66 (16.33)	▲0.68 (15.38)	0.79 (2.88)	▲0.66 (13.37)	▲0.74 (3.86)	▲0.76 (4.26)
B			0.61 (2.61)	0.66 (3.00)	0.70 (2.67)	0.71 (1.09)	0.69 (1.96)	0.64 (0.31)	0.69 (0.36)
C				0.83 (0.00)	0.81 (0.07)	▲0.80 (12.25)	0.73 (0.18)	▲0.70 (6.00)	▲0.70 (6.00)
D					0.84 (0.08)	▲0.75 (9.80)	0.80 (0.25)	▲0.78 (8.00)	▲0.78 (8.00)
E						▲0.78 (8.05)	0.76 (0.05)	▲0.76 (6.37)	▲0.84 (9.31)
F							▲0.80 (9.00)	▲0.75 (5.00)	▲0.80 (6.25)
G								0.75 (0.20)	0.80 (0.25)
H									0.88 (0.00)

表 3. 一致率

▲は有意差があることを示す

(McNemar 検定、α =0.05)。

上. 医学専門用語か否か

グループ内の平均は、GⅠが 0.73,
GⅡが 0.79, GⅢが 0.81。

下. 難解語か否か

グループ内の平均は、GⅠが 0.69,
GⅡが 0.80, GⅢが 0.86。

	GⅠ(医師)		GⅡ(医療関係者)				GⅢ(一般)		
	A	B	C	D	E	F	G	H	I
A		▲0.69 (25.00)	0.86 (0.82)	0.81 (0.07)	▲0.79 (17.00)	▲0.79 (17.00)	▲0.79 (17.00)	▲0.75 (20.00)	▲0.76 (15.21)
B			▲0.63 (26.13)	▲0.68 (22.15)	▲0.90 (8.00)	▲0.88 (6.40)	▲0.83 (4.57)	0.86 (2.27)	▲0.88 (6.40)
C				0.88 (1.60)	▲0.73 (18.18)	▲0.73 (18.18)	▲0.73 (18.18)	▲0.66 (19.59)	▲0.73 (18.18)
D					▲0.78 (14.22)	▲0.78 (14.22)	▲0.80 (16.00)	▲0.71 (15.70)	▲0.78 (14.22)
E						0.93 (0.00)	0.90 (0.00)	0.86 (0.82)	0.90 (0.00)
F							0.90 (0.00)	0.86 (0.82)	0.90 (0.00)
G								0.84 (0.69)	0.85 (0.00)
H									0.89 (1.00)