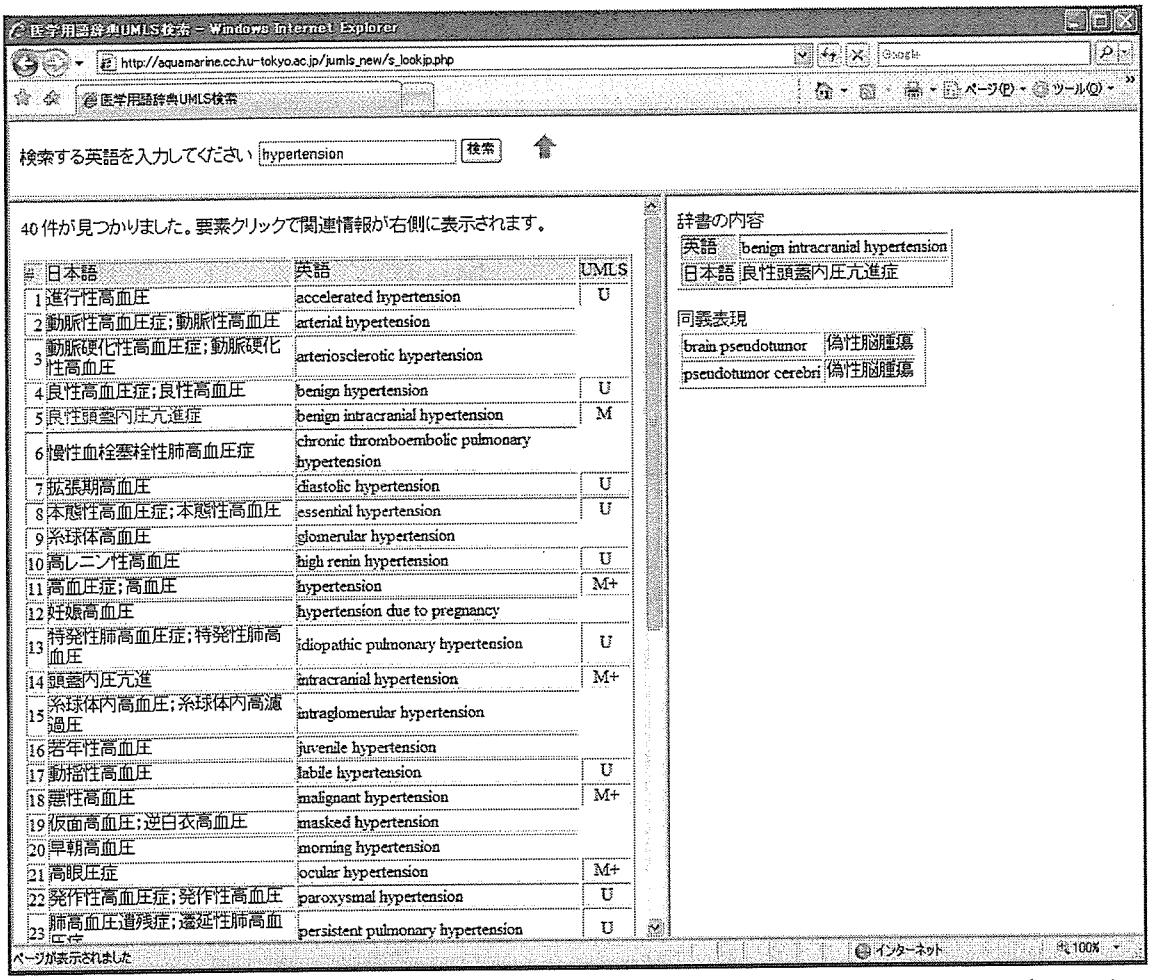


図 1d. 英語の検索語から辞書の英語を検索し、左画面に辞書の内容がリストされる。各要素をクリックすることで付加情報を参照する点は日本語の場合と同様である。

（4） 日本語医学用語からの MEDLINE 検索

システムの概念図を図 2 に示す。日本語医学用語を検索語として与えることにより、医学用語辞典の中でその部分文字列を含み、かつ MeSH と対応の取れている語彙のリストが表示される。この語彙の中から MEDLINE で検索したいものをクリックすると、画面右側にその語彙に対応する MeSH term を Pubmed で検索した画面が表示される。

また、共出現検索を行うこともできる。共出現検索の欄にチェックを入れて、上記と同様に検索を行って表示された語彙リストからいずれかの語彙をクリックすると、その語彙に共出現した MeSH term（ひとつの文献上で共起した MeSH term）に対応する日本語医学用語が共起頻度順にリストされる。例えばブロンズ糖尿病という医学用語（MeSH での代表語はヘモクロマトーシス）に共出現する MeSH term は頻度順に膜タンパク質（302 件）、クラス I 抗原（257 件）、鉄（121 件）、…、となる。そのリストの中から検索したい語彙、例えば「鉄」をクリックすると、「ブロンズ糖尿病」と「鉄」の両方を含む文献が Pubmed で検索され、その結果が画面右側に表示される。この様子を図 3a-b に示す。

なお、NCBI には、Pubmed の他に多くのデータベースがあり、それらを横断的に検索する画面も用意されている。そこでその横断検索に直接アクセスするインターフェイスを図 3c に示す。PubMed の他にも検索を行いたい場合には有用である。

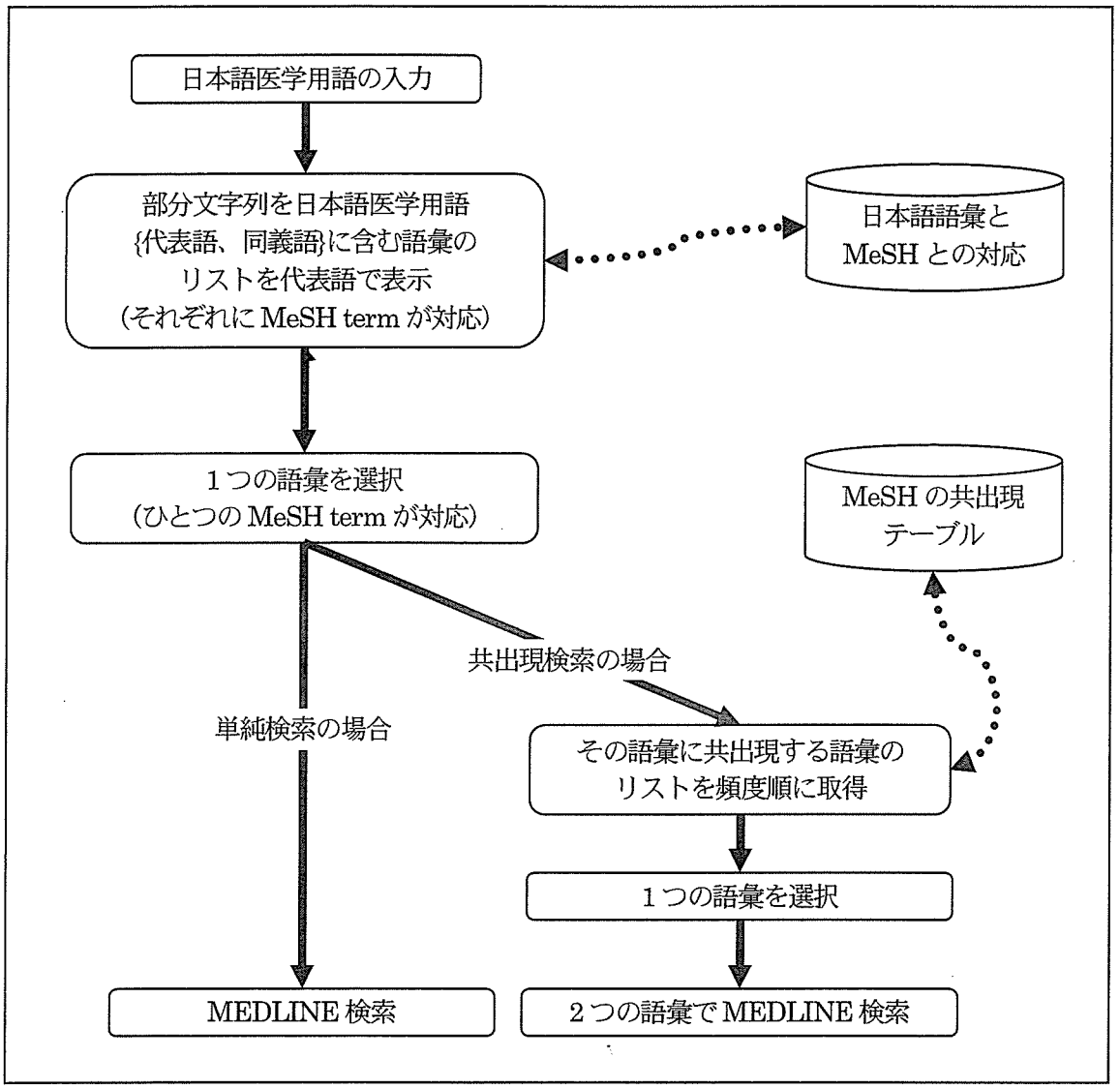


図2.　日本語によるMEDLINE検索システムの処理図

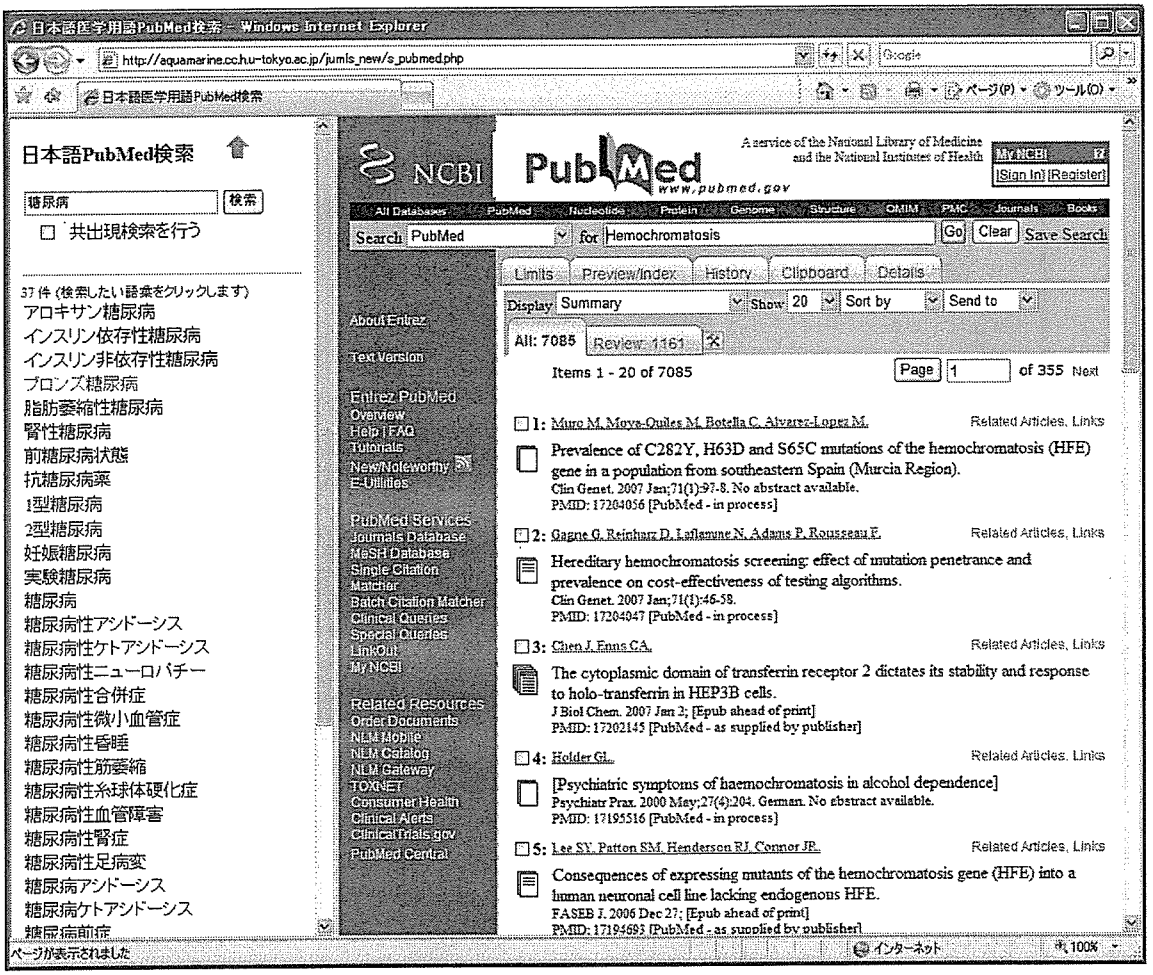


図 3a. 日本語の検索語から辞書の英語を検索し、左画面に検索された辞書の日本語がリストされる。この中の語彙をクリックすることにより、PubMed 検索が行われ、結果が右画面に表示される。

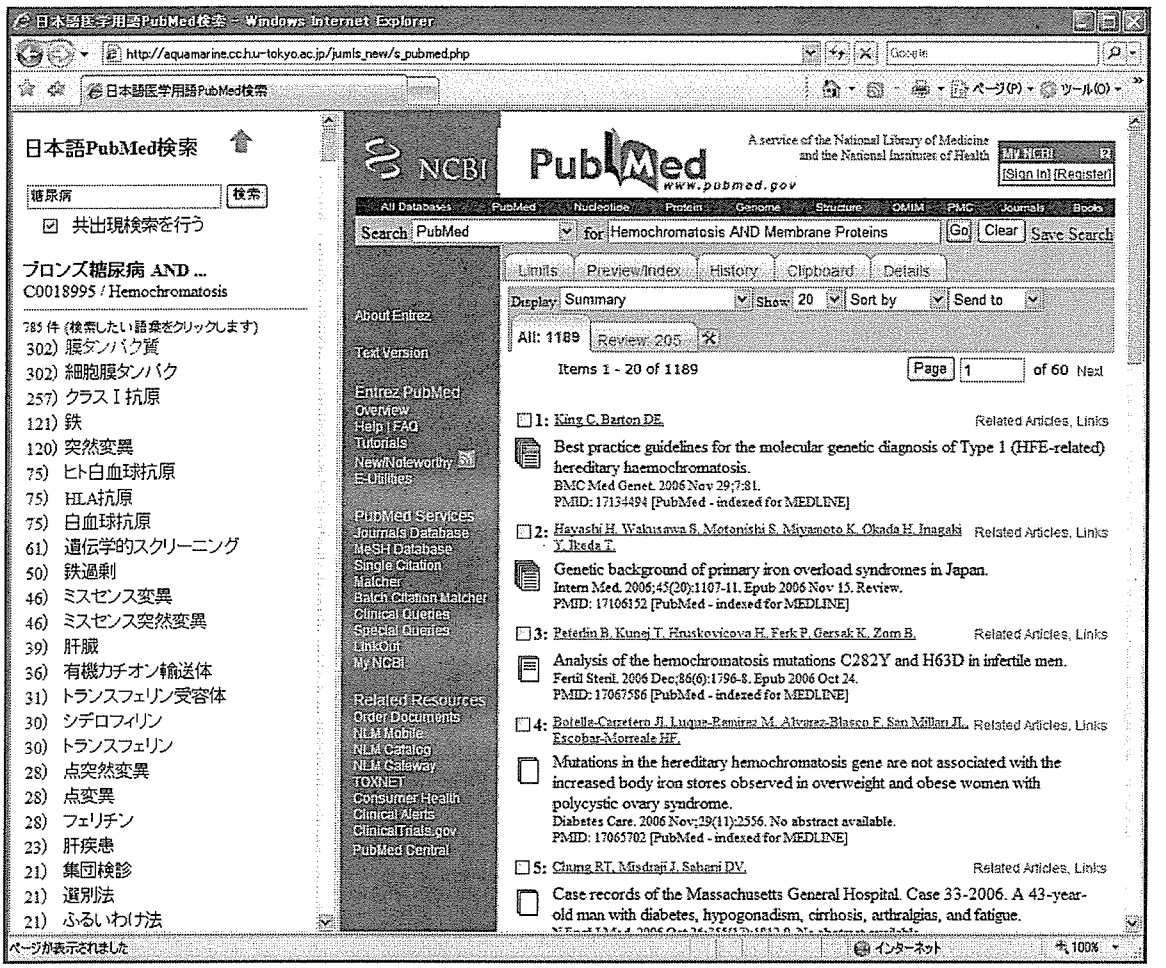


図 3b. 共出現検索欄をクリックして検索を行うと、図 3a と同じリストが画面左に表示される。そこでいずれかの語彙をクリックすると、その語彙に共出現する頻度の高い順に語彙がリストされる。ここでは「ブロンズ糖尿病」と共起する語彙が頻度順に 785 件表示されている。各語彙の左に示されている数字は、その共出現数である。この中の語彙をクリックすることにより、ふたつの語彙を両方とも含む文献が PubMed で検索され、結果が右画面に表示される。

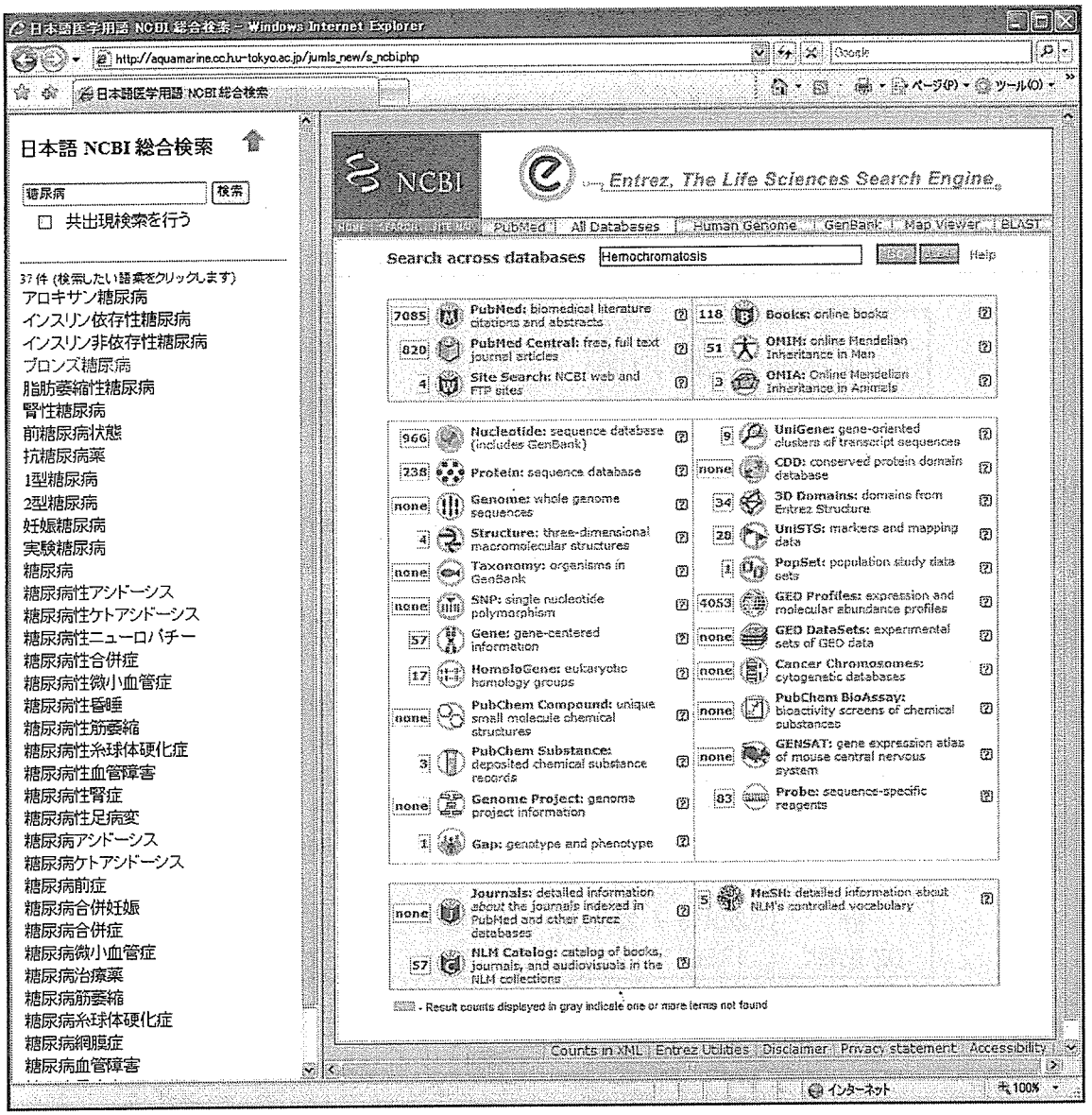


図 3c. PubMed 検索と同じく、NCBI の横断検索をおこなった画面。インターフェイスは PubMed 検索と同様であるが、右画面には NCBI の横断検索の結果が表示される。

（5） 評価 （旧システムとの比較による新システムの評価）

新システムと旧システムで、検索に利用できる MeSH の概念数、対応する日本語の概念数と表記の数、および日本語概念に含まれる表記の平均を調べた結果を表 2 に、また MeSH の概念カバー率（システムで検索に使うことのできる語彙に対応した MeSH 概念の、全 MeSH 概念に占める割合）を MeSH のカテゴリごとに比較した結果を表 3 に示す。

また旧システムの日本語（異なり数 45,706）と新システムで MeSH と対応している日本語（異なり数 24,578）を比較したところ、総和は 57,471、共通は 12,813 であり、旧システムにあって新システムにない日本語は 32,893、新システムにあって旧システムにない日本語は 11,765 であった。これを図 4 に示す。

表2.　日本語からMEDLINE検索を行うシステムの評価指標（新旧システムの比較）

	旧システム	新システム
対応MeSH概念数/全概念数	18,919 / 20,743 (91.2%)	17,813 / 23,885 (74.5%)
対応する日本語表記の数	45,706	25,478
1概念あたりの日本語表記数	2.42	1.43

表3　MeSHカテゴリごとの概念のカバー率
（カバー率とは、検索に利用できるMeSH概念と全MeSH概念との比）
（なお2002ACにはカテゴリVが存在していないため、それ以外を提示）

	旧システム	カバー率	新システム	カバー率
A	1,686 / 1,704	98.9%	1,848 / 1,895	97.5%
B	1,699 / 2,278	74.6%	2,371 / 2,594	91.4%
C	6,385 / 6,496	98.3%	6,750 / 6,817	99.0%
D	8,879 / 9,155	97.0%	4,127 / 10,395	39.7%
E	2,237 / 2,331	96.0%	2,527 / 2,554	98.9%
F	845 / 867	97.5%	736 / 758	97.1%
G	2,046 / 2,110	97.0%	2,315 / 2,385	97.1%
H	465 / 484	96.1%	507 / 522	97.1%
I	377 / 450	83.8%	427 / 461	92.6%
J	235 / 253	92.9%	266 / 298	89.3%
K	106 / 151	70.2%	146 / 176	83.0%
L	311 / 321	96.9%	324 / 362	89.5%
M	134 / 173	77.5%	174 / 187	93.0%
N	1,039 / 1,265	82.1%	1,267 / 1,322	95.8%
Z	1 / 368	0.3%	1 / 371	0.3%
計	26,445 / 28,406	93.1%	21,941 / 29,225	75.1%
DZを除く	17,565 / 18,883	93.0%	19,658 / 20,331	96.7%

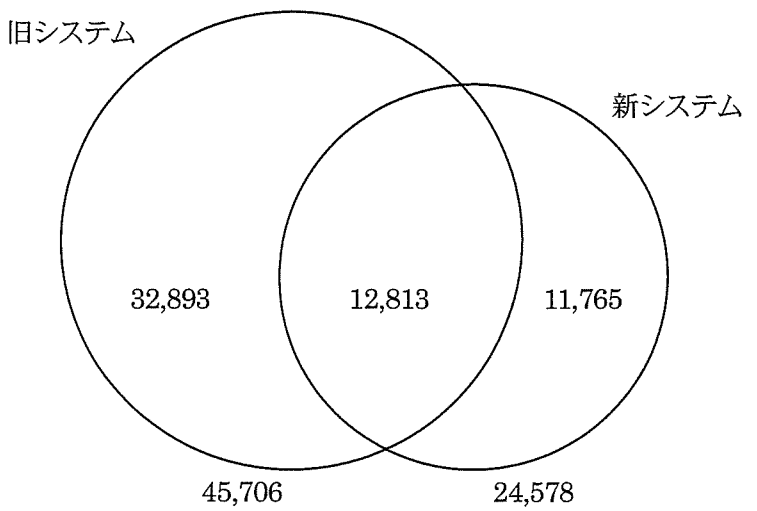


図4．新旧システムにおけるMeSHと対応する日本語語彙の異なり数。
旧システムと比べて新システムでは語彙数が約半分であるにもかかわらず、その中のおよそ半数は旧システムにはない語彙である。

4．1．4　考察

システムの評価として新旧システムの比較を行った表2の結果を見ると、本研究で開発した新システムはMeSH概念のカバー率も、MeSHに対応する日本語の数も、旧システムに劣るように見える。しかし表3でカテゴリごとに細かく見ると、その内容は旧システムとは大きく異なっていることがわかる。DやZカテゴリの語彙は医学用語辞典には積極的には収載されていないので、これらのカバー率は低くなっているが、他のカテゴリでは全体的に見て旧システムよりもカバー率が高くなっていることがわかる。また図5に示したように、新システムの日本語異なり語彙数は旧システムの約半分であるにもかかわらず、その中のおよそ半分の語彙は旧システムには存在しない。この結果と表3の結果とを合わせて考えると、新システムでは検索に有効な新規語彙を十分に備えた上でカテゴリごとにバランス良く語彙が選択されていることを示していると言える。特にMeSHの全概念数が2002AC（20,743）から2006AC（23,885）になって増加していることを考え合わせると、実質的に医学で利用される一般的な用語を、本システムでは良好にカバーしていると言える。

以上により、旧システムは国内で刊行された医学文献の索引付けに用いられている用語集を元にしているため、文献検索には相性が良いものの、日常臨床で使われる用語とはやや性質が異なることが懸念されていたわけであるが、それが医学用語辞典の語彙を利用することにより、大きく改善されたことを示すものと考えられる。新システムは使い勝手という面では未評価であるが、少なくとも検索可能性（つまりより多くの日本語語彙がMeSHと対応付けられていること）という観点からは、医学用語辞典を出発点とした本研究のMEDLINE検索システムは有用である可能性が高いと結論できる。

ただし、MeSHの1概念あたりの日本語表記数は明らかに新システムの方が少なくなっており、このこと自体は検索に不利であることは否めない。これは医学用語辞典の性質、すなわち異なる表記の中から日本医学会として推奨する表記を選択する、という特徴を持った辞典であることから仕方のないことであるが、検索システムとして利用する場合には同義語あるいは異表記をもっと広範囲にリンクする仕掛けが必要であると考えられる。

また本研究では日本医学会医学用語辞典の英語見出しの中から正規化によってMeSHと一致した語彙を選択しているにとどまっている。この作業は人手で行われたものであるために信頼性は高いが、継続的なメンテナンスを行うことは困難であると考えられる。今後はMeSHに対応していない語彙との関連性情報などまでを含めた包括的な管理、および正規化による一致率を高めるアルゴリズム上の工夫、それによる対応作業の自動化などが望まれる。また医学用語辞典が各医学専門領域の用語集を統合したものであることに踏み込むならば、医学専門領域ごとのオントロジーを開発・維持・管理した上で、それらをマージする、その際に必ずしもMeSHに捕らわれることなく、MeSHとの対応関係を保った上で、オントロジーの構造を優先させるという方法も考えられるが、いずれも今後の課題である。

なお、現在はNLMでも他種類の言語を使ってPubMedを検索することのできるシステムであるBabelMeSH（Multilanguage Search for MEDLINE/PubMed (Beta)）が試作されている（http://babelmesh.nlm.nih.gov/）。これは英語だけではなくアラビア語、中国語、フランス語、ドイツ語、イタリア語、日本語、ポルトガル語、ロシア語、スペイン語の医学用語によりMEDLINEを検索することを可能とするシステムである。このシステムでは2004年からUMLSに加わった日本語MeSHを元にしていると推測される。しかし本研究で示されたように、MeSHの日本語訳や当初より文献検索という目的のために体系付けられた用語集を利用した検索システムには、検索を行うユーザーに親和性の高い用語であるとは限らない点で、一定の限界もあると考えられる。文献索引用語だけにとらわれることなく、本研究のように日常臨床で使用する日本語とMeSHとの関連を確立することが、一般の人々にとってより使いやすい文献検索システムの開発に有用であることをコメントしておきたい。

参考文献

[1] Coletti MH, Bleich HL. Medical subject headings used to search the biomedical literature. J Am Med Inform Assoc. 8(4):317-323, 2001.

[2] Pritchard SJ, Weightman AL. MEDLINE in the UK: pioneering the past, present and future. Health information and libraries journal. 22 Suppl 138-44, 2005.

[3] Sood A, Ghosh AK. Literature search using PubMed: an essential tool for practicing evidence-based medicine. The Journal of the Association of Physicians of India. 54303-308, 2006.

[4] Hersh WR, Donohoe LC. SAPHIRE International: a tool for cross-language information retrieval. Proceedings / AMIA Annual Symposium. 673-677, 1998.

[5] Boyer C, Baujard V, et al. HONselect: multilingual assistant search engine operated by a concept-based interface system to decentralized heterogeneous sources. Medinfo. 10(Pt 1):309-313, 2001.

[6] Onogi Y, Ohe K, et al. Mapping Japanese medical terms to UMLS Metathesaurus. Medinfo. 11(Pt 1):406-410, 2004.

[7] Nelson SJ, Schopen M, et al. The MeSH translation maintenance system: structure, interface design, and implementation. Medinfo. 11(Pt 1):67-69, 2004.

[8] Lindberg DA, Humphreys BL, et al. The Unified Medical Language System. Methods of information in medicine. 32(4):281-291, 1993.

[9] Humphreys BL, Lindberg DA, et al. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc. 5(1):1-11, 1998.

表4. 医学用語辞典の見出し語がMeSHと一致したものに関して、MeSHカテゴリごとの概念数、およびMeSH全体の概念数、そして医学用語辞典におけるMeSH概念の採用率（カテゴリの上位3桁分類で集計）

カテゴリ	分類	辞典	MeSH全体	カバー率
A01	身体	75	75	100.0%
A02	筋骨格系	192	193	99.5%
A03	消化器系	71	73	97.3%
A04	呼吸器系	41	41	100.0%
A05	泌尿生殖器系	78	79	98.7%
A06	内分泌系	40	42	95.2%
A07	心臓血管系	117	117	100.0%
A08	神経系	299	301	99.3%
A09	感覚器	93	95	97.9%
A10	組織	114	118	96.6%
A11	細胞	382	411	92.9%
A12	体液と分泌物	59	61	96.7%
A13	動物の身体構造	54	55	98.2%
A14	口顎系	77	78	98.7%
A15	血液免疫系	86	86	100.0%
A16	胚構造	56	56	100.0%
A17	外皮系	14	14	100.0%
B01	動物	783	857	91.4%
B02	藻類	48	51	94.1%
B03	細菌	523	698	74.9%
B04	ウイルス	493	499	98.8%
B05	真菌	138	142	97.2%
B06	植物	1134	1145	99.0%
B07	古細菌	35	59	59.3%
C01	細菌性疾患と真菌性疾患	326	327	99.7%
C02	ウイルス性疾患	226	226	100.0%
C03	寄生虫疾患	139	140	99.3%
C04	腫瘍	643	649	99.1%
C05	筋骨格系疾患	283	287	98.6%
C06	消化器疾患	287	287	100.0%
C07	口顎疾患	191	195	97.9%
C08	気道疾患	189	189	100.0%
C09	耳鼻咽喉疾患	98	98	100.0%
C10	神経系疾患	749	757	98.9%
C11	眼疾患	203	203	100.0%
C12	泌尿器疾患と男性生殖器疾患	168	170	98.8%
C13	女性生殖器疾患と妊娠合併症	183	186	98.4%
C14	心臓血管疾患	337	340	99.1%
C15	血液疾患とリンパ疾患	237	239	99.2%
C16	先天性,遺伝性および新生児疾患と奇形	538	546	98.5%
C17	皮膚疾患と結合組織疾患	325	326	99.7%
C18	栄養疾患と代謝性疾患	295	298	99.0%

C19	内分泌系疾患	144	148	97.3%
C20	免疫系疾患	204	205	99.5%
C21	環境起因障害	283	286	99.0%
C22	動物の疾病	121	128	94.5%
C23	病理学的状態，症状，徴候	581	587	99.0%
D01	無機化学物質	284	425	66.8%
D02	有機化学物質	522	1776	29.4%
D03	複素環式化合物	302	1080	28.0%
D04	多環式化合物	162	478	33.9%
D05	高分子物質	85	179	47.5%
D06	ホルモン，代用物質，拮抗剤	124	197	62.9%
D08	酵素と補酵素	413	1327	31.1%
D09	炭水化物	195	357	54.6%
D10	脂質	139	256	54.3%
D12	アミノ酸、ペプチド、タンパク	856	2546	33.6%
D13	核酸など	110	287	38.3%
D20	複合混合物類	116	183	63.4%
D23	皮膚作用剤	308	651	47.3%
D25	医用材料と歯科材料	68	117	58.1%
D26	薬物	49	61	80.3%
D27	化学作用と効用	394	475	82.9%
E01	診断	529	535	98.9%
E02	治療	413	415	99.5%
E03	麻酔法と無痛法	34	35	97.1%
E04	外科手術	401	407	98.5%
E05	研究及び検査手法	735	744	98.8%
E06	歯科学	177	178	99.4%
E07	機器と資材用品	238	240	99.2%
F01	行動と行動メカニズム	350	363	96.4%
F02	心理現象と過程	223	230	97.0%
F03	精神障害	163	165	98.8%
F04	行動訓練と活動	152	154	98.7%
G01	生物科学	93	93	100.0%
G02	保健医療業務	183	183	100.0%
G03	環境と公衆衛生	397	412	96.4%
G04	生物学的現象，細胞生理学，免疫	276	280	98.6%
G05	遺伝学的過程	110	112	98.2%
G06	生化学的現象，代謝，栄養	251	257	97.7%
G07	生理学的過程	137	140	97.9%
G08	生殖と泌尿生殖器の生理学	146	150	97.3%
G09	循環生理学と呼吸生理学	142	143	99.3%
G10	消化器系生理学，口腔生理学，皮膚生理学	45	46	97.8%
G11	筋骨格系生理学，神経系生理学，眼生理学	166	169	98.2%
G12	化学的現象と薬理学的現象	52	55	94.5%
G13	遺伝学的現象	86	87	98.9%
G14	遺伝学的構造	231	258	89.5%
H01	自然科学	507	522	97.1%

I01	社会科学	301	321	93.8%
I02	教育	86	88	97.7%
I03	人間活動	40	52	76.9%
J01	技術，産業，農業	168	190	88.4%
J02	食品と飲料	98	108	90.7%
K01	人文科学	146	176	83.0%
L01	情報科学	324	362	89.5%
M01	人間	174	187	93.0%
N01	人口集団特性	91	92	98.9%
N02	保健医療の施設，人的要員，サービス	371	380	97.6%
N03	保健医療経済と保健医療組織	299	322	92.9%
N04	保健医療サービス管理	264	277	95.3%
N05	保健医療の品質，アクセス，評価	242	251	96.4%
V01	出版タイプ	3	23	13.0%
Z01	地理的位置	1	371	0.3%

4．2　MeSH 翻訳管理システムの医学用語シソーラスへの適用（MTMS：MeSH Translation Maintenance System）

4．2．1　研究趣旨

MTMS を使用した MeSH 翻訳語の提供に関する取り組みと経緯、及び今後の展開を検討した。

（1）　研究目的

米国国立医学図書館(National Library of Medicine:以下　NLM)が医学文献検索に使用しているメタシソーラス、UMLS（Unified Medical Language System）は多種の医学用語を包括しており、多言語にも対応が進んでいる。2003 年までに、フランス、ドイツ、イタリア、オランダ、フィンランド、ポルトガル、ロシア、スペインの 8 カ国語に渡り、Concept レベルでの体系化ができており、スウェーデン、ポーランドと言った東欧ばかりでなく、ベトナムを始めとするアジアばかりでなく、漢字文化の日本、韓国の対応も進行中である旨が発表されており、多言語対応のシソーラスとしても、幅広く展開されてきた。NLM は UMLS の多言語対応をより良く、取り込めるシステムとして MeSH Translation Maintenance System(以下、MTMS)を開発した。今回 NLM より、医学中央雑誌刊行会に日本語医学用語の取り込みに限って提供された MTMS を使用して、日本語医学用語をいかに提供し、メンテナンスを行っていくかを試みた。（図 1、図 2）

（2）　研究方法

英語圏以外の国々でも Web による情報検索サービス、PubMed が幅広く利用されていることは言うまでもない。だが、自国の言語で PubMed 等の検索に利用できる医学用語辞書、あるいはシソーラスと言ったものを作成しているところは、先にも述べたとおり、僅か 8 ヶ国にすぎない(2001 年当時)。英語圏ではない東欧のチェコなども、MeSH の翻訳に取り組んでいるが、UMLS から見れば数パーセントの対応である。これは先にも述べた 8 ヶ国共々対応が完全に取れているとは言えないのが実状である。ところが、医学中央雑誌刊行会が提供している「医学用語シソーラス」は 99%の用語が MeSH とリンクが取れており、検索ばかりでなく、論文等で使用されている英語の医学用語も幅広く取り入れており、なおかつ MeSH と Concept レベルでのリンクがある。このことは UMLS へ日本語医学用語の提供をある程度容易にすることができた原因でもある。

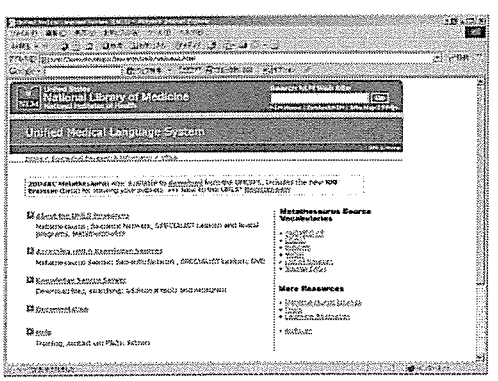


図 1　UMLS のトップ画面

医学中央雑誌刊行会発行の「医学用語シソーラス」から、MeSH に対応する日本語の用語を抽出し、MTMS にてデータベースを構築し、登録された用語に対して、各々、更新、修正していく。それらの用語が NLM の Supervisor の手によって承認されると、UMLS へ登録される。この手順を必ず踏むことで UMLS への日本語医学用語が登録されていく。

図 3 に示した画面から、用語を入力し、検索し、その用語に関してそれぞれ更新等の作業を行う。

検索された用語は Concept ごとに Entry Term、Entry Term(Non Print)等に分けて表示される（図 4)。NLM による承認などの情報はフォント及び色によって表現される。

新規登録の場合は対象の MeSH 用語を検索してレコードを呼び出し、必要項目を入力していく。当然のことながら日本語入力(自国語入力)が可能である。(図 5) それぞれのメニューがあり、呼び出された画面に対して更新、修正等を行っていく。また、登録された内容をカテゴリー構造で表示することも可能である。

入力が自国語で可能なことは、大きな価値がある。副産物ではあるが、このシステムを利用することによって、MeSH の変更（毎年、数回、カテゴリー等の変更が行われている）に合わせて、当会のシソーラスのメンテナンス性も高まり、互いに良いものを提供し、構築することが可能になっている。

MeSH 部門のチーフである、Dr. Stuart Nelson によれば、2003 年に NLM で行われた、Board of Regent において、日本を皮切りに、漢字圏の用語集も UMLS に取り入れていくというプランが発表されたとのことである。

他の例としては、ベトナムがこのプランに勢力的に動いており、中国も本年度から、伝統医学に限った TCM-MeSH（Traditional Chinese Medicine-MeSH）が UMLS に採用されると、中国中医科学院：CACMS(Chinese Academy of Chinese Medical Science)の Prof. Fan から、昨年 12 月に聞いている。中国では、独自の CUMLS(Chinese Unified Medical Language System)などを開発しており、Dr. Cui も中国独自の MeSH を開発し、私どもが取り組み医学中央雑誌刊行会が行っている作業も、今後は取り組んでいく考えを持っていると思われる。伝統医学を保健医療に用いている韓国は、どちらかというと、関心が薄い状態である。

漢字圏のみならず、英語圏以外の用語の UMLS への取り組みは今後、様々な形で行われていくであろう。

図２　トップ画面

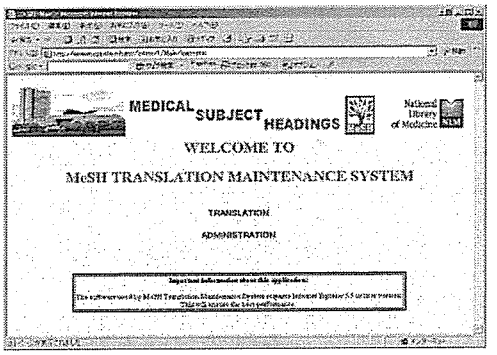
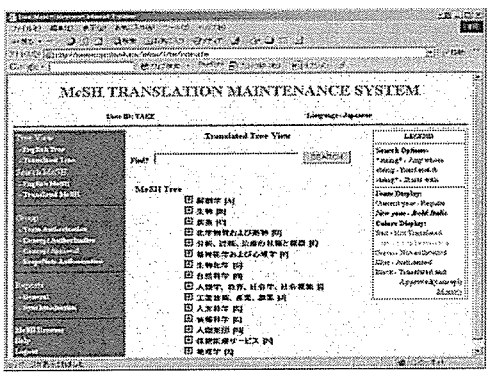


図３　Translated MeSH のトップ画面

図４　レコードの表示



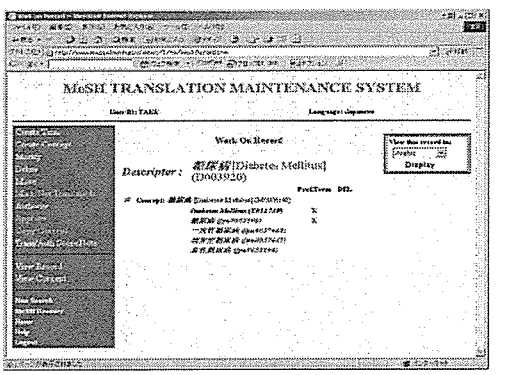
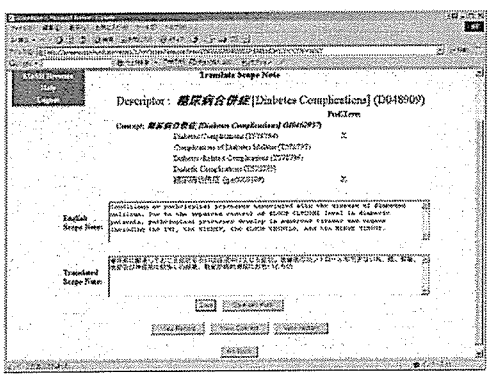


図５　ScopeNote の入力



（３）研究結果

MTMS による用語のメンテナンスに関しては十分に行える。今回、約２万語の用語を登録し、MeSH にて変更のあったものに関しては更新作業を行ったが、問題なく実行することができた。

（4） 考察

今回の作業はUMLSの中でもMeSHに関してのみではあるものの、行うことが可能であった。同時に、他の言語に対しても、対応することが十分に可能であることが示唆された。

（5） 結論

今回の作業が他の言語への対応するテストケースとしては十分に適応でき得ることがわかった。この成果は情報検索の支援、索引作業の支援ばかりでなく、各種、英語圏以外のデータベースへのリンクも十分に可能な手法であると思われる。

別紙４

研究成果の刊行に関する一覧表レイアウト

書籍

著者氏名	論文タイトル名	書籍全体の編集者名	書　籍　名	出版社名	出版地	出版年	ページ

雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
Yuzo Onogi	Assigning Categorical Information to Japanese Medical Terms Using MeSH and MEDLINE	Proceedings of the 11th World Congress on Medical Informatics		(in print)	2007

Assigning Categorical Information to Japanese Medical Terms Using MeSH and MEDLINE

Yuzo Onogi[a], MD, PhD

[a] Clinical Bioinformatics Research Unit, Graduate School of Medicine, the University of Tokyo, Japan;

Abstract

*This paper reports on the assigning of MeSH (Medical Subject Headings) categories to Japanese terms in an English-Japanese dictionary using the titles and abstracts of articles indexed in MEDLINE. In a previous study, 30,000 of 80,000 terms in the dictionary were mapped to MeSH terms by normalized comparison. It was reasoned that if the remaining dictionary terms appeared in MEDLINE-indexed articles that are indexed using MeSH terms, then relevancies between the dictionary terms and MeSH terms could be calculated, and thus MeSH categories assigned. This study compares two approaches for calculating the weight matrix. One is the TF*IDF method and the other uses the inner product of two weight matrices. About 20,000 additional dictionary terms were identified in MEDLINE-indexed articles published between 2000 and 2004. The precision and recall of these algorithms were evaluated separately for MeSH terms and non-MeSH terms. Unfortunately, the precision and recall of the algorithms was not good, but this method will help with manual assignment of MeSH categories to dictionary terms.*

Keywords: Medical Subject Headings, MEDLINE, Algorithms, Classification, Japanese medical terms

Introduction

Clinical documents are processed and stored in electronic format by many providers in Japan and elsewhere, and this trend will certainly increase in the future. Some clinical data is stored as numbers and codes, for example laboratory examination results and disease names or categories, but the remainder is stored as natural text, for example data relating to a patient's present illness, clinical findings, and radiology reports. If we want to make use of this accumulated information or drive any decision-support system based on the clinical documents, then natural text processing is indispensable, which is very difficult for Japanese-language documents.

Japanese is an agglutinative language, meaning that there are no separators such as spaces between words. There are computer tools for morphological analysis [1] that can segment a whole sentence into parts of speech, and can derive syntactic information. If these tools are used for medical texts, it is necessary for medical terms to be added to the vocabulary, because the above-mentioned tools were originally made for general, non-technical, purposes. However, it is difficult to proceed further with deep parsing because there are very few resources on semantics for Japanese medical terms. Here we define semantic information as relating to the category or categories that a term belongs to. For example, "gallbladder" (in Japanese) is the name of an anatomical organ, and "gallbladder calculus" is the name of a clinical finding.

To obtain such categorical information, medical taxonomies are useful, because terms are located within hierarchies of categories, although the categories vary between taxonomies. One such taxonomy is the Japanese medical thesaurus published by the Japan Medical Abstract Society (JAMAS), in which terms are mapped to MeSH (Medical Subject Headings) terms. Furthermore, Japanese MeSH terms have been released by the National Library of Medicine (NLM) since 2002. We verified the validity of the mapping between the JAMAS thesaurus and MeSH terms, so we were able to obtain categorical information for about 20,000 concepts and about 50,000 Japanese medical terms [2].

An English-Japanese dictionary published by the Japan Medical Society (JAMS) comprises about 80,000 entries, and is authored by medical subcommittees within Japan. This dictionary is an authority for standardized medical terminology in Japan. To determine what proportion of this dictionary is covered by MeSH, we compared terms in the dictionary with those in the Unified Medical Language System (UMLS) [3-5] by normalization (norm and lvg programs in the Lexical tools of the UMLS) and found that about 30,000 terms can be mapped to MeSH terms, but that the remaining 50,000 terms cannot be mapped to any existing controlled vocabularies in UMLS, and therefore we cannot obtain any categorical information for these 50,000 terms [6]. We reasoned that if these unmatched terms appeared in the titles and abstracts of articles indexed by MEDLINE, then we could calculate the most relevant MeSH category for the term,

because each article is indexed using MeSH terms. Thus, the objective of this study was to assign categorical information to Japanese terms that appear in an English-Japanese medical dictionary using MEDLINE and MeSH.

Methods

The basic premise of this study is that if terms in the above-mentioned English-Japanese dictionary appear in the titles and abstracts of articles indexed by MEDLINE, and that if each article is indexed by MeSH, then a list of relevancies between dictionary terms and MeSH categories could be calculated, so the most precise categorical information for a term in the dictionary could be obtained by selecting the largest (most important) weight category for the term in the list of MeSH categories.

Materials used in this study were as follows: the 2005AA version of the UMLS published by the NLM; the 2005 version of the MEDLINE records leased from the MEDLARS Management Section of the NLM; and the English-Japanese medical dictionary published by the Japan Medical Association, which includes about 80,000 entries in English.

Terms in the dictionary were compared one at a time with titles and abstracts of articles indexed by MEDLINE, with both converted to lower-case letters, for articles published between 2000 and 2004 (a total of 5 years), and the frequencies (number of appearances of each term per article) were obtained. The comparison of terms was performed directly; that is, neither stemming nor inflection conversion was carried out.

Relevancies between dictionary terms and articles were obtained by the TF*IDF weighting method [7] as follows:

$$w_{i,j} = tf_{i,j} \times idf_i = \frac{freq_{i,j}}{\arg\max_l freq_{l,j}} \times \log\frac{N}{n_i} \qquad (1)$$

where i represents a term; and j represents a document; w is a weight value for term i in document j; $tf_{i,j}$ is a "term frequency," that is, the frequency of term i in document j, and it is normalized against the number of times the most frequent term l appears in the document j; idf_i is the "inverse document frequency" for term i, that is, the reciprocal number of documents including term i (n_i), which is also normalized against the total number of documents, N.

Similarly, relevancies between articles and MeSH terms (main headings) were calculated. Here we used both major and minor MeSH terms indexed to an article.

We obtained two weight matrices: one comprised dictionary terms versus articles, and the other comprised articles versus MeSH terms. The inner products of these matrices gave us a weight matrix of relevancies between dictionary terms and MeSH terms as follows:

$$\begin{bmatrix} entry\ terms \\ articles \end{bmatrix} \bullet \begin{bmatrix} articles \\ MeSH\ terms \end{bmatrix} = \begin{bmatrix} entry\ terms \\ MeSH\ terms \end{bmatrix} \qquad (2)$$

This method is hereafter referred to as the "proposed method." In order to evaluate the performance of this method, we calculated another weight matrix between dictionary terms and MeSH terms directly from all combinations of dictionary terms and MeSH terms for an article (referred to as the "baseline method") and compared the results.

Evaluations were performed by calculating precision and recall for both algorithms. We constructed two kinds of test sets (gold standards), one consisting of MeSH terms (main headings or synonyms by themselves) existing in the dictionary (9770 terms in total), and the other consisting of non-MeSH terms randomly chosen from the dictionary and categorized manually (100 terms) by a researcher other than the one who developed the algorithms. For each dictionary term, we obtained a list of categories and corresponding weight values in descending order, which were smaller if the category was less relevant. Then precision and recall for rank n was defined as follows:

$$precision_n = \frac{N(retrieved_{rank \le n} \wedge correct)}{N(retrieved_{rank \le n})} \qquad (3)$$

$$recall_n = \frac{N(retrieved_{rank \le n} \wedge correct)}{N(test\ set)} \qquad (4)$$

where N indicates the number of terms. ROC (Receiver Operating Characteristic Curve) can be drawn using a series of precision and recall values for various n values.

This algorithm should not be dependent on the features of the dictionary terms, but if the frequencies were calculated from MEDLINE-indexed articles only, then there might be no difference between the performances of the MeSH and non-MeSH terms. We wanted to investigate this possibility further.

Results

Identification of dictionary terms in MEDLINE

The total number of articles indexed in MEDLINE between 2000 and 2004 (a 5-year period) was about 2.8 million. The number of dictionary terms found in the titles and abstracts of these articles was 49,384 (62% of terms in the dictionary). Of these, 22,296 (28%) were non-MeSH terms; that is, we assigned for the first time some categorical information to these dictionary terms (Table 1). In addition, we compared

the dictionary terms with the titles and abstracts of articles indexed in MEDLINE between 1995 and 2004 (a 10-year period), and found that 53,670 terms appeared, of which 25,227 (31%) were non-MeSH terms. This shows that the number of dictionary terms appearing in MEDLINE seems to reach a plateau at around a 5-year period.

We were able to calculate two kinds of weight matrices (dictionary terms versus articles and articles versus MeSH terms) for the 10-year period, but it was impossible to calculate the inner products of the matrices because of the limitations of our computer resources.

Table 1- Number of dictionary terms in the titles and abstracts of articles indexed in MEDLINE over 5-year and 10-year periods.

	No. of terms (%)
Total no. of terms in the dictionary	80,131 (100)
No. of matched with MeSH terms	33,487 (42)
Five-year period	
No. appearing in MEDLINE	49,384 (62)
No. non-MeSH terms	22,296 (28)
No. unassigned terms	24,348 (30)
Ten-year period	
No. appearing in MEDLINE	53,670 (67)
No. non-MeSH terms	25,227 (31)
No. unassigned terms	21,417 (27)

Evaluation of categorization algorithms

The precision and recall results for the baseline and proposed methods are shown in Figure 1. This shows that the proposed method has better precision and recall (0.83 and 0.78, respectively) for the first rank result in both test sets compared with the baseline method (0.71 and 0.71, respectively). However, the performance results for non-MeSH terms were poor, with the proposed method showing better precision and recall (0.51 and 0.56, respectively) than the baseline method (0.26 and 0.26 respectively).

Table 2 shows a sample result of the categorization of non-MeSH terms in the dictionary using the proposed method. Japanese terms and their corresponding English terms are presented, and the English terms are categorized by using the proposed method to the "Assigned" categories. When this category correctly matched the given "Answer" category, then the symbol in the "Correct" column is "O," otherwise it is "X."

Discussion

Matching of dictionary terms with MEDLINE

Finding the dictionary terms in the titles and abstracts of articles indexed in MEDLINE was partially successful, because we were able to find about 50,000 terms in the present study. However, about half of these were MeSH terms, so we were only able to find 23,000 non-MeSH terms. The number of terms found in MEDLINE using articles indexed during a 10-year period was not significantly greater than that from a 5-year period, and so the number seems to reach a plateau after 5 years. This finding implies that we should use resources other than articles to find the dictionary terms. For the dictionary terms that didn't appear in MEDLINE, we attempted to find the reasons why this had occurred. Three main reasons were apparent. Firstly, we found that most were certainly valid medical terms, but were seldom used in recent articles. They comprised such terms as names of bacteria, infectious diseases, physiological tests or manipulations, and signs and symptoms. Secondly, there were many Latin terms, which also seldom appear in recent articles. Based on these analyses, it may be better to use older articles in MEDLINE to efficiently find these old terms. Thirdly, we found some terms that may be in common use in English written by Japanese-speakers, but which may not be commonly used by authors from other countries. Next we checked the validity of terms not found in MEDLINE using MerckSource (http://www.mercksource.com/) to determine whether they had corresponding definitions or whether other term was suggested as a substitution, because the system could not find any corresponding terms and supposed to be misspelled. We found approximately 4,000 terms that were specific to Japan (e.g. relating to Japanese legislation) but invalid terms comprised a majority of the unassigned terms. If these invalid terms can be removed, then more terms could be matched with MEDLINE in the future.

We compared the dictionary terms and MEDLINE text directly in the present study, and this may have reduced the number of terms matched. Stemming (or canonical normalization in lvg) may be effective to increase the number of matches, but may be ineffective to calculate relevancies because false matches will increase when using stemming.

We used only titles and abstracts of articles indexed by MEDLINE in the present study, but there is a possibility we could increase the number of matches if we could use the entire text of articles, because the more terms that are compared, the more matches will be found not only with respect to number but also with respect to variations.

Categorization algorithm

The proposed method – using the inner products of two weight matrixes (dictionary terms versus articles and articles versus MeSH terms) – seems to perform better than the baseline method (dictionary terms versus MeSH terms directly) when the category that has the most significant weight value for each term is selected. However, the baseline method was slightly better when using third and lower ranks. For non-MeSH terms, the proposed method performed better than the baseline for all ranks, although precision and recall were both about 0.5 for the first ranked weight value. This result was not good, with only half of the non-MeSH terms being categorized accurately. The purpose of this study was to

assign categorical information to Japanese dictionary terms, and the results so far show that automatically categorizing terms with this algorithm fails to assign terms correctly. However, this does not imply that the output of the proposed method is useless, rather we think that this assignment will help in the manual determination of which category a term belongs to (because only the first-ranked category was chosen in the proposed method). For this purpose, the proposed method, using the inner products of two weight matrices, seems to be suitable.

We expected that there would be no difference in the performance of these algorithms for MeSH terms and non-MeSH terms but, as shown in Figure 1, there were actually differences. We checked the non-MeSH terms and found that the average document frequency was about half that of the MeSH terms, but that the standard deviations were almost the same. This shows that the non-MeSH terms do not frequently appear, and that the TF*IDF weight value mainly depends on document frequency. However, some non-MeSH terms have very large document frequencies (because we did not eliminate stop words and non-MeSH terms contain stop words such as At - astatine), causing the poor result, which was different from the results for MeSH terms.

The nature of the relevancy data should also be discussed. We obtained relevancies between dictionary terms and MeSH terms, but actually this implies collocation or co-occurrence, rather than the two having a categorization or subsumption relationship as we expected. For example, the term "hypertension" may have stronger relevancy to the term "angiotensin-converting enzyme" (drug category) than it does to "vascular diseases" (disease category) as categorized in MeSH. Furthermore, we cannot identify the meaning of each relevancy based on the TF*IDF method only. However, as is shown by the precision for MeSH terms, subsumption might happen to be the dominant relationship in the relevancies determined in the present study.

Conclusion

In order to obtain categorical information for Japanese terms in a English-Japanese medical dictionary, we calculated weight matrixes between English dictionary terms and MeSH terms with the TF*IDF method using MEDLINE title and abstract data for articles published between 2000 and 2004. To calculate the matrix, we used the inner products of two weight matrices. We found 22,000 non-MeSH dictionary terms and assigned corresponding MeSH categories. Although precision and recall was not good, the results are still useful for our purposes.

Acknowledgements

This research was supported in part by funding for a project entitled "Application of Japanese medical thesaurus linked to UMLS Metathesaurus" funded by Japan's Ministry of Health, Labor and Welfare.

References

[1] Kurohashi S, Nagao M, Japanese Morphological Analysis System JUMAN version 3.5. Department of Informatics, Kyoto University. (in Japanese); 1998.

[2] Onogi Y, et al. Mapping Japanese Medical Terms to UMLS Metathesaurus. Proceedings of the 11th World Congress on Medical Informatics, 2004: pp. 406-410.

[3] Lindberg D, Humphreys B, McCray A. The Unified Medical Language System. Methods Inf Med 1993: 32(4); pp. 281-91.

[4] Campbell KE, Oliver DE, et al. Representing Thoughts, Words, and Things in the UMLS. JAMIA. 1998: 5; pp. 421-31.

[5] Cimino JJ. Auditing the Unified Medical Language System with Semantic Methods. JAMIA. 1998: 5; pp. 42-51.

[6] Onogi Y. Knowledge support system for medical information retrieval in Japanese using UMLS Metathesaurus. Proceedings/Japan Congress of Medical Informatics 2004; pp. 386-7

[7] Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing & Management 1988: 24(5); pp. 513–523

Address for correspondence

Yuzo Onogi, MD, PhD: **yonogi@hcc.h.u-tokyo.ac.jp**
Clinical Bioinformatics Unit, University of Tokyo Hospital, 7-3-1 Hongo Bunkyo Tokyo, 113-8655 Japan