別添 1

厚生労働科学研究費補助金

医療技術評価総合研究事業

ＵＭＬＳと連携する日本語医学用語シソーラスの実用性に関する評価研究

平成１７年度　総括・分担研究報告書

主任研究者　脊山　洋右

平成１８（２００６）年４月

別添 2

目次

別紙３

厚生労働科学研究費補助金　（医療技術評価総合研究事業）
総括分担研究報告書

UMLS と連携する日本語医学用語シソーラスの実用性に関する評価研究

主任研究者　脊山洋右　お茶の水女子大学生活科学科教授
分担研究者　鈴木博道　(財)国際医学情報センター主任研究員

研究要旨

平成 16 年度までの研究成果を、日本医学会医学用語辞典英和第 3 版改訂作業に応用し、これまでの成果の再検討、再評価をすることを目的として、辞典改訂のために英語医学用語の自動的なチェック、自動分類、自動計算による判定補助、を実験した。ボキャブラリー・ファイルが文献検索用のものであったことから、文献検索用の用語についてはこの手法がそのまま有効であった。しかし、医学用語辞典の場合には SNOMED などの様なボキャブラリー・ファイルを活用した方がより有効であることが示唆された。いずれにしても、方法論的には有効かつ実用性が裏付けられた。

分担研究者一覧

開原成允　国際医療福祉大学副学長
野添篤毅　愛知淑徳大学教授
小野木雄三　東京大学大学院助教授
篠原恒樹　医学中央雑誌刊行会理事長
鈴木博道　(財)国際医学情報センター

A．研究目的

平成 17 年度最大の課題は、これまでに築いて試みてきているシソーラス開発手法が日本医学会医学用語辞典英和第 3 版改訂において、有効かつ適切に利用できるものであるかどうか、利用できるとするとどの部分でどの様に利用可能であるかをテストすることである。広い意味では、方法論の再検討・再確認であると言えよう。

また同時に、これまでの研究活動を継続し、ＵＭＬＳへの日本語搭載を推進、国際共同研究プロジェクトのＵＭＬＳに参画しているＮＬＭや他の研究者との情報交換・交流も推進する。

B．研究方法

日本医学会医学用語辞典の見出し語を正規化処理しＵＭＬＳにマッチングを行うことで、医学用語辞典見出し語の検証をした。また、シソーラス化された際の実用性を把握するために、検索時にシソーラスがどの様に機能すべきか PubMed を中心に検討を加えた。実用化後の課題として、シソーラスの更新・改訂についての検討も行った。

NLM 関係者、UMLS 研究者との定期的な情報交換を継続すると共に、本研究班の成果発表も兼ねたシンポジウム「医学用語・シソーラスを巡る国際的展開――UMLS そして日本語医学用語シソーラス」を 3 月 2 日に開催した。このシンポジウムでは、別途招へいしたオレゴン健康科学大学教授の William Hersh 博士による “Controlled Terminologies in Biomedicine: Rationale, Challenges, and Limitations” と題する講演も行った。

（倫理面への配慮）本研究は患者情報を扱うものではなく、医学用語のみを扱っており、倫理面への配慮は不要と考えている。

C．研究結果

日本医学会医学用語辞典の英語医学用語 8 万

語に対して正規化マッチングをした結果、40%がマッチングに成功した。マッチしない語については、スペルミス、表記上の不一致の他、日本医学会医学用語辞典では形容詞も見出し語となっていることなど、原因を分析した。

スペルミスと覚しきマッチしなかった語について、UMLS 並びに各種インターネット上の情報源を利用してスペルミスのチェックをかけ、4,500 語について人的なチェックと修整を経て、修整を完了した。人的チェックを容易にするため、UMLS を利用して分野毎に自動的に振り分けを行った。

シソーラスのメンテナンスを配慮して、NLM で開発した MTMS (MeSH Translation Maintenance System) を調査し、UMLS、MeSH、医学用語シソーラス、医学用語辞典などでテスト使用した。

D．考察

スペルミスを機械的に検出し人的にちぇっくしやすい様に振り分けまでする行程は有効であった。このための自動計算のロジックはさほど複雑なものでは無く、有効性を占う課題はむしろ材料となったボキャブラリー・ファイルの特性によることが明らかとなった。今回手をつけることが出来なかった SNOMED 利用は、MeSH による手法より医学用語辞典にはふさわしいものと思われる。

UMLS を１つの標準として、英語医学用語を点検し整理する機械的手法は応用可能であったが、日本語の医学用語についてどの様に適用するか、今後の課題であろう。

E．結論

日本医学会の医学用語辞典第 2 版の英語見出し語を材料としたスペルミスの検出、UMLS と正規化により一致する語彙の特定、それ以外の語彙を適切な専門分野に振り分けて判定を効率的に行う手法を開発し、た。また MeSH に一致した語彙の同義語として、MeSH ではなく UMLS から同義語を収集する手法を実験した。医学用語辞典に存在する表記を MeSH に準拠した概念に関連づけ、同時に既存辞典に存在しない同義語を新たに導入することが可能となった。

手法としては確立したものの、文献検索のための MeSH と医学用語辞典の用語とでは違いが大きく、適切な材料を設定することが緊要であろう。

G．研究発表

1. 論文発表

なし。

2. 学会発表

小野木雄三. UMLS と Medline を利用した日本語医学用語への意味属性付与. 医療情報学, 25(suppl.) 933-936, 2005

H．知的財産権の出願・登録状況（予定を含む。）

1. 特許取得

なし。

2. 実用新案登録

なし。

3. その他

なし。

シンポジウム （平成17年3月2日開催）

医学用語・シソーラスを巡る国際的展開
－UMLS そして日本語医学用語シソーラス－

私共は平成 13 年度以来、厚生労働省医療技術評価総合研究の研究費補助金を受け、日本語医学用語シソーラス開発手法の開発研究とその試行を行ってきております。本研究は米国国立医学図書館ＮＬＭ (National Library of Medicine)との協力によりＵＭＬＳ (Unified Medical Language System) との連携に基礎をおいております。この度、米国から医学用語やＵＭＬＳに造詣の深い Oregon Health & Science University の William Hersh 教授をお招きし、米国内での状況をご講演頂き、併せて、本研究班の成果の一端を広くご紹介する機会を持つことを企画しました。

医学用語や医学シソーラス、医学にとらわれることなく広く専門用語管理やシソーラスにご興味有る方々、シソーラスなどを通じて医学医療情報の流通に関わり有る方々など、多くの方々のご参加をお待ちしております。

平成 17 年度厚生労働科学研究「UMLS と連携した日本語医学用語シソーラスの実用性に関する評価」研究班　（主任研究者　お茶の水女子大学教授　脊山洋右）

プログラム

15:00～15:30	あいさつと研究概要　　脊山洋右（お茶の水女子大学）
15:30～16:30	特別講演（英語） "Controlled Terminologies in Biomedicine: Rationale, Challenges, and Limitations" William Hersh, M.D. Professor and Chair, Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University
16:30～17:00	質疑応答など 進行　　小野木雄三（東京大学大学院）
17:00～	研究班員・協力者などとのフリーディスカッション

日時

日時:　2006年3月2日（水）15:00～17:00
会場:　第6セミナー室 （東京大学医学部新研究棟 13 階）
参加費:　無料
共催:　特定非営利活動法人 医学中央雑誌刊行会
　　財団法人 国際医学情報センター

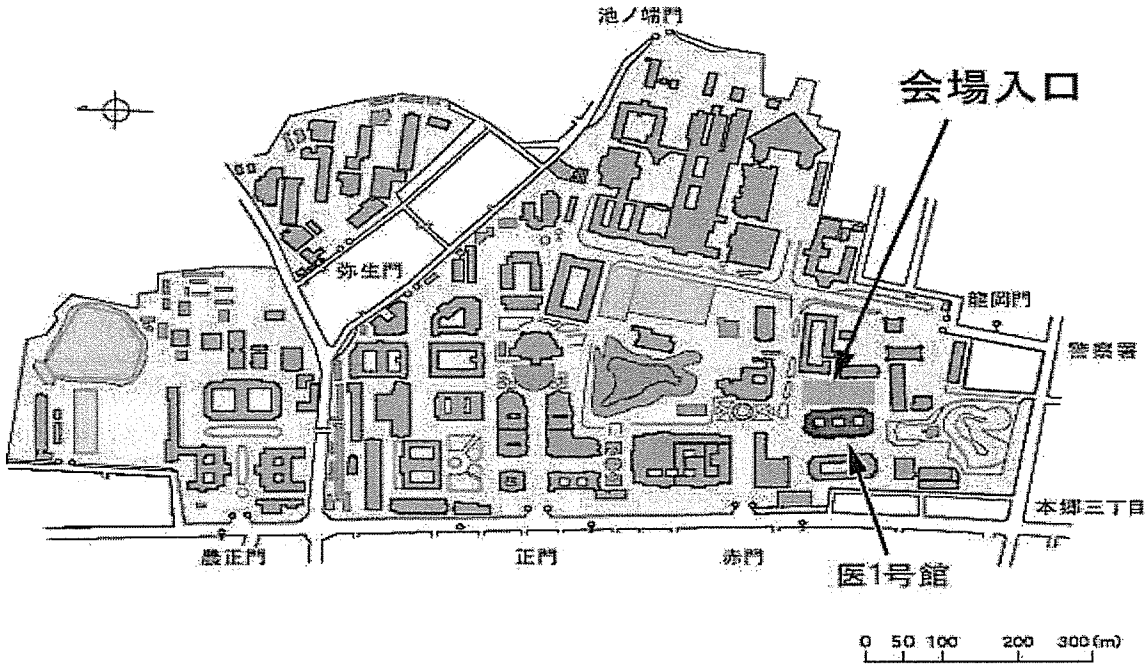
参加申し込み方法

参加ご希望の方は、所属機関名・会社名、所属部署 ・役職、氏名（ふりがな）、電話番号 E-mail、をご記入の上 e-mail または faxにてお申し込み下さい。特に受講票は発行致しませんので、当日、そのまま会場にお越し下さい。

－ お問い合せ先 －
（財）国際医学情報センター
事業推進室 UMLS シンポジウム担当
〒160-0016 東京都新宿区信濃町 35
信濃町煉瓦館 3 階
TEL 03-5361-7089　FAX 03-5361-7110
E-mail jigyo@imic.or.jp

会場：第６セミナー室（東京大学医学部新研究棟 13 階）

本郷キャンパス施設案内図



http://www.biochem2.m.u-tokyo.ac.jp/web/shimizu/kinkyou020314.html

きりとり線

お申し込みは E-mail jigyo@imic.or.jp または FAX 03-5361-7110

機関名 会社名			
所属部署 役　職			
ふりがな			事務使用欄 受付番号
氏　名			
住　所	〒		
連絡先	TEL	FAX	UMLS WS 2006.3.2
	E-mail		

厚生労働科学研究費補助金 (医療技術評価総合研究事業)
分担研究報告書

UMLS と連携する日本語医学用語シソーラスの実用性に関する評価研究
各種コードとの連携

分担研究者　開原　成允　国際医療福祉大学大学院長

研究要旨

現在、日本における最も権威のある医学用語集としては、日本医学会の編纂した医学用語辞典があるが、この辞典には二つの大きな欠点があった。第一は、この医学用語辞典に採用されている英語が必ずしも英語圏での標準的な英語医学用語ではないこと、第二は、シソーラスの概念がないために、基幹となる用語も同義語も同列に扱われ、利用者にどの用語が基幹的な用語であるかがわからないことであった。

本研究では、この欠点を改善するために、UMLS の Concept Unique Identifier (CUI)を用いて、医学用語辞典の英語と UMLS の連携を図ると共に、日本の用語辞典にもシソーラスの概念を定着させることを目的とした。

研究方法としては、医学用語辞典の英語と UMLS とを正規化した上でマッチングをとったが、約 8 万語の中の 40%でマッチングがとれた。また、逆に MeSH に存在しているが、医学用語辞典に存在していない語があった。また、厚生労働省の標準病名集の日本語病名との間の整合性も検討した。

キーワード: 医学用語、UMLS、シソーラス、標準病名集

A．研究目的

現在、日本における最も権威のある医学用語集としては、日本医学会の編纂した医学用語辞典がある。この辞典は、英語の医学用語を日本語に変換する際の標準的な日本語医学用語として使われてきたが、二つの大きな欠点があった。第一は、この医学用語辞典に採用されている英語が必ずしも英語圏での標準的な英語医学用語ではないこと、第二は、シソーラスの概念がないために、基幹となる用語も同義語も同列に扱われ、利用者にどの用語が基幹的な用語であるかがわからないことであった。

本研究では、この欠点を改善するために、UMLS の Concept Unique Identifier (CUI)を用いて、医学用語辞典の英語と UMLS の連携を図ると共に、英語のシソーラスを医学用語辞典に埋め込むことを研究する。また、第二は、対応する日本語については、医学会各分科会の用語辞典を参照しつつ、基幹となる日本語と同義語を区別することを研究する。

この方法が確立した後は、これを実際の医学用語辞典の改定に役立てることが本研究の目的である。

B．研究方法

日本医学会医学用語辞典・英和の電子版を用い、その英語を正規化し、UMLS の CUI の英語との間でマッチングをとる。マッチ（対応がとれた）した英語は、それが基幹的英語か同義語かを区別して医学用語辞典の英語に対応するコード番号を付す。これによって、まず、医学用語辞典の中の国際的にみて標準的な英語を区別する。また、採用した基幹となる語は、「推奨語 (preferred term)」と呼ぶこととした。

次に、マッチしなかった英語について、それが重要な基幹語であるか、または削除してもいいものかを判定し、現在の英語単語を整理する。

次にマッチした英語に対応する日本語について、1 対 1 の対応であれば、それを日本語の推奨語として採用する。もし、複数の日本語が対応している場合には、医学会各分科会の用語辞典を参照しつつその中の推奨語と同義語を判別し、その印を付す。また、病名については、厚生労働省と医療情報システム開発センターが開発したいわゆる「標準病名集」があり、広く普及しはじめている。この病名集は日本医学会の監修を経たものであり、推奨語と同義語が整理されている。従って、病名についてはこの病名集との間の整合性をとることによって、日本医学会用語集を整理することを試みる。

C．研究結果

現在、日本医学会医学用語辞典には、83791 の英語が収載されている。単純に対応をとると、この中の約40%がUMLSとの対応をとることができた。

しかし、対応がとれたものの中にも、UMLS のさまざまなレベルでの対応関係があるので、それを個別に精査し、英語見出し語として採用する語を確定した。　一方、UMLS の中で重要な位置を占める MeSH に存在する英語でありながら、医学用語辞典に存在していない見出し語がある。これについては、それを精査し、医薬品などの名称や国名など索引には必要であるが、辞書としては必要のない語を除いて、追加すべき英語見出し語を確定した。

さらに、医学用語辞典の見出し英語の中で UMLS に対応つかなかった英単語については、さまざまな原因が考えられた。第一は、UMLS は名詞のみであるが、日本医学会用語辞典は形容詞をも含むからこれは対応がとれない。また、単純なスペルミスがまだ残っているものもあった。また、薬剤やアイソトープなど表記上の問題が原因のものもあった。これについては、その原因別に検討を開始したが作業は平成 17 年度には終了しなかった。

複数の日本語訳を持つものの整理については、日本医学会の各分科会の協力を得ることとし、各分科会に要請した。

標準病名集との間の整合性についての検討も本年はまだ実現していないが、漢字の略字体の問題、カナ表記の必要な用語にカタカナを用いるか、ひらがなを用いるかなど、多くの解決しなければならない問題が残されている。

D．考察

これまで、多くの医学用語集が作られたきたが、通常はさまざまな場所で使われている英語を無差別に収集して出発点とした用語集も多く、そこに収載されている英語が国際的に標準的な英単語であるか否かについてはほとんど検討されたことはなかった。今回 UMLS を一つの基準として、収載する英単語を整理するという試みはこれまでにないものである。

また、これまでの医学用語集にはシソーラスの概念がまったく存在せず、これも大きな混乱の原因となっていた。今回の研究によって日本にもシソーラスの概念が定着すれば非常に大きな意義がある。

E．結論

本年は、第二年目として、日本医学会医学用語辞典の見出し英単語と UMLS の CUI の間の対応を研究し、多くの作業が進展した。

これらの結果は、医学用語辞典の改訂に反映される予定である。

しかし、英単語の整理は可能となっても、日本語医学用語の整理は行えない。この部分は日本の医学会の総力をあげて行う他はない。平成 17 年度には、各分科会との協力体制が確立した。

G．研究発表

1. 論文発表
   なし。
2. 学会発表
   なし。

H．知的財産権の出願・登録状況
（予定を含む。）

1. 特許取得

   なし。

2. 実用新案登録

   なし。

3. その他

   なし。

厚生労働科学研究費補助金 （医療技術評価総合研究事業）
分担研究報告書

UMLS と連携する日本語医学用語シソーラスの実用性に関する評価研究
UMLS との機械的連携に関する研究

分担研究者　小野木雄三　東京大学大学院医学系研究科
クリニカルバイオインフォマティクスユニット臨床情報工学部門

研究要旨

医学用語とは医学に関する何らかの概念を文字列で表記したものである。一般に、ひとつの概念には多くの表記が存在し（同義語）、逆にひとつの表記には多くの概念が対応する（多義語）。これは用語の混乱を招くと同時に医療文書の電子的処理を障害する要因となっている。従来の医学用語辞典にはこの観点が欠けていたため、代表語と同義語の区別、および MeSH®（文献を索引付けするための統制用語集）との対応が検討されている。そのために本研究では辞書の英語見出し語のスペルミスの検出、UMLS®(Unified Medical Language System)と一致する語彙の特定、一致しない語彙に対しては専門分野への振り分けを行った。また MeSH の同義語を辞書の同義語に含める手法を検討した。

A．研究目的

表記と概念との対応には同義性と多義性を考慮する必要がある。日本医学会の医学用語辞典に代表語と同義語の区別、MeSH との対応を付加するために、①辞書の英語見出し語のスペルミス検出、②UMLS と一致する語彙の特定、③一致しない語彙の専門分野への振り分けを行う。次に④MeSH 同義語を辞書の同義語に含める際に整合性の取れる手法を検討する。

B．研究方法

日本医学会編纂による医学用語辞典第 2 版 3 刷英和版に収載されている英語見出し語約 8 万語を対象とした。

①スペルミス検出：UMLS(2005AA) と MerckSource や Google などインターネット上のリソースを利用してスペルチェックを行い、UMLS に一致するものと語彙の定義が得られたものは信頼できる語彙として除外、それ以外を誤り候補として抽出し、専門家による修正を行った。

②UMLS および MeSH との一致：見出し語彙からストップワードを空白に置換し、小文字に変換し、得られた単語を整列する(正規化)処理により、UMLS に収載されている語彙と比較を行った。また空白とハイフンに対しては、それらを削除した場合も検索した。UMLS には MeSH 日本語版も含まれているが、まだ信頼性に欠けると判断され、日本語文字列の比較は適用しなかった。

③専門分野への振り分け：UMLS 中に一致するものが存在しない語彙に対し、どの分野の専門家に判定を依頼すれば良いかをある程度自動的に判定することを課題とした。専門分野として MeSH のカテゴリーを利用し、その語彙が MEDLINE のタイトル・抄録の過去 5 年分の中に出現したか否か、出現したならばどの MeSH で索引付けされているか、を調べて TF×IDF 法により語彙と MeSH カテゴリーとの関連性を計算し、同時に文献中の語彙出現数を測定した。別に 100 例の正解を作成し、適合率と再現率を測定した。

④MeSH の同義語と辞書同義語との違い：MeSH は文献を索引付けする目的で構築された用語集であるため、概念の粒度が粗いものがある。例えばひとつの MeSH 概念が、辞書側の複数の(同義ではない)語彙に対応する。つまり MeSH の同義関係をそのまま辞書の同義関係に利用することはできない。これを解決するために MeSH より粒度の細かい UMLS の概念を利用した。

（倫理面への配慮）本研究は患者情報を扱うものではなく、公開された医学用語辞典と文献情報を利用するものであるため、倫理面への配慮を要することはない。

C．研究結果

①スペルミス検出により、約 4500 語の辞書語彙を特定し、専門家による修正を行うことができた。次に②正規化処理により UMLS および MeSH と一致した語彙は約 34000 語、空白やハイフンの削除で一致した追加分は 600 語程度であった。残りの約 42000 語のうち、過去 5 年間の MEDLINE 中に出現した語彙は約 23000 語であり、残り約 19000 語は使用頻度の少ない語彙と考えられた。③この 23000 語に対して振り分け処理を行い、語彙に関連する MeSH カテゴリーを関連性の強い順に 10 個まで計算した。別に用意した 100 語彙に対し、1 位にランクされたカテゴリーの適合率と再現率はそれぞれ 0.5 程度、2 位までを含むと再現率 0.6、適合率 0.4 であった。④最後に辞書語彙の同義語として「一致した MeSH 概念に含まれる同義語を採用する」のではなく「辞書語彙と一致した MeSH 語彙に割り当てられている UMLS 概念コードを有する MeSH 語彙の同義語を採用する」ことを提案した。

D．考察

スペルミスの検出ではインターネット上の使用頻度はあまり参考にならなかった。これは検索数が 0 であった場合以外に適切な閾値を設定できないことによる。しかし、MEDLINE 文献中に存在しない語彙と併せ、インターネット上での使用頻度によって、辞書見出しに使用頻度を考慮した順位付けを付す可能性が開けたと考える。

語彙の専門分野への自動振り分けの成績は良好とは言えないが、判定作業を支援する目的には叶うものと考えられる。今後も関連づけの手法や MeSH 以外のカテゴリーへの割り付けなどを検討したい。

同義語の採用方法では UMLS 概念の粒度がある程度は細かいことを仮定している。実際 22995 個の MeSH 概念中、2 つ以上の UMLS 概念が対応するものの数は 7647 個であり、この方法はある程度適切であり、少なくとも MeSH 語彙をそのまま辞書の同義語と考えるよりもはるかに有用であると考えられる。しかし SNOMED-CT の概念粒度は UMLS 概念よりも細かいことが知られており、辞書語彙の概念粒度がどのあたりに位置するかについては今後検討する必要がある。

E．結論

日本医学会の医学用語辞典第 2 版の英語見出しに対してスペルミスの検出、UMLS と正規化により一致する語彙の特定、それ以外の語彙を適切な専門分野に振り分けて判定を効率的に行う手法を開発した。また MeSH に一致した語彙の同義語として、MeSH ではなく UMLS から同義語を収集する手法を提案した。以上により、医学用語辞典に存在する表記を MeSH に準拠した概念に関連づけ、同時に既存辞典に存在しない同義語を新たに導入することが可能となった。

G．研究発表

1. 論文発表

なし。

2. 学会発表

小野木雄三. UMLS と Medline を利用した日本語医学用語への意味属性付与. 医療情報学, 25(suppl.), in print, 2005

H．知的財産権の出願・登録状況
（予定を含む。）

1. 特許取得

   なし。

2. 実用新案登録

   なし。

3. その他

   なし。

厚生労働科学研究費補助金（医療技術評価総合研究事業）
分担研究報告書

UMLSと連携する日本語医学用語シソーラスの実用性に関する評価研究
医学用語標準化に向けて
UMLS、MeSHから見た情報標準化への取り組み

分担研究者　篠原　恒樹　NPO医学中央雑誌刊行会 理事長
豊玉　速人　NPO医学中央雑誌刊行会 システム管理課

研究要旨
医学文献検索を効率的に行うためには、辞書・シソーラスを活用し、有効な検索用語を利用できるようにすることが必要である。有効な検索用語をいかに構築し、提供するか、UMLSとMeshと医学用語シソーラス収載用語を中心として、検討した。

A．研究目的
医学文献検索において、効率よく検索結果を導くには、有効な検索用語の選択が必須である。米国国立医学図書館(National Library of Medicine：以下、NLM)が提供しているPubMed/MEDLINEにはMeSHのようなシソーラスばかりではなく、多言語対応及び他の用語集にも対応しているUMLSのようなメタシソーラスも存在する。これらの有効的活用法を探求していくと、日々変化すると言っても過言ではない用語の変化に対応しなければならないことは言うまでもない。今回、我々は有効な検索用語をいかに構築し、提供する方途を現存する辞書に収載されている用語をNPO医学中央雑誌刊行会（以下、医中誌）が発行している「医学用語シソーラス　第5版」に収載されている用語と比較検討し、医学文献の検索に対応する有効な用語集作成方法を検討した。

B．研究対象と結果
国内で発刊されている医学用語辞書の中から、日本医学会が発行する「日本医学会用語集・英和　第2版」を対象とした。「日本医学会用語集・英和　第2版」は2000年に改定されたが、以降、一部の修正等を除き大幅な改定作業は行われていない。MeSH、UMLSの改定に比べるとかなりの年月が経ている。MeSHは1年に1度、UMLSは年3回の追加、変更等が行われている。医中誌が発行する「医学用語シソーラス」もこの間改定作業が行われ、第5版を発行している（「日本医学会用語集　第2版発行時は第4版であった」。この間にはかなりの変遷があることが予測される。標準的に使用されている辞書に掲載されている用語を対象に検索に使用される用語との変化を追跡してみる。辞書とシソーラスは用途が違い、形態そのものも違うが、ユーザーからみれば論文記述の際、検索目的を考慮しながら執筆している著者も数多くいると思われる。以前、当会では「日本生化学会」の予稿集の分類作業を「日本生化学会」が演題募集の際に使用している用語集と医中誌が発行している「医学用語シソーラス」を比較検討し、演題分類作業を試験的に行った経験がある。その際にわかったことは、①標題や内容に関係なく、付与されているキーワードがかなり存在する。要するに検索目的とは違う意味でのキーワード付与が投稿者によって行われている。②実際に分類作業を行う作業者は著者が付与されたキーワードをかなりの割合で考慮しなければならない。③実際には②のようなキーワードはあまり意味をなさない場合が多い。これらの結果から分類作業者は用語の変化は分類作業者の経験値にかかる比重がかなり大きい。実際には検索結果に近いものを要求されながらも多少違うと思われるものを作成しているケースが多く存在している。著者が付与するキーワードは著者の意図が入って

おり、この差に我々は注目し、検討してみたところ、投稿者、分類作業者、双方の意識の違いはそれぞれの意識する用語の変遷の違いと言っても良いことがわかった。

C. 結論

「日本医学会用語集　第 2 版」の中には約 70,000 の用語が収載されている。2005 年に NLM から発行された UMLS2005AA(2005 年初版:AA、AB、AC と年 3 回買改定されている）に収載されている MeSH を中心とした用語と比較検討したところ、約 42,000(60%)の用語が UMLS には存在しない英語であった。うち約 20,000 の用語は文部科学省が発行した「科学技術用語集・医学編」に、約 18,000 語が「英辞朗」のような英和辞書に収載されている用語であったが、出典が不明であるものが約 4,000、存在した。これらを詳細に見てみると、あきらかなスペルミスなどもあったが、ほとんどが現存しない英語表現とも言うべき用語であった。

約 28,000 の用語が MeSH になんらかの形で存在しているものであった。①文字列が完全に一致、②複数のディスクリプタと一致したもの、③MeSH の SubHeadings（副標目）あるいは Supplementary Concept Record（薬物、化学物質）に存在する用語と一致したもの、④直接は MeSH の用語と一致しないが、UMLS の他の用語から Concept レベルでの一致がみれれるもの、以上の 4 種類が存在した。①は 1,500 語、②は 500 語、③は 3,000 語、④は 15,000 語という内訳であった。①に関しては現存する有効な検索用語である。②は単数形、複数形と言った違いもあるが、辞書とシソーラスの違いを考慮すれば容易に分けられた。③は化学物質、薬物は辞書という観点から見るとなくてもよいと思われるものもあるが、概ね問題ないと思われる。④に関しては、文字列から言ってもまったく違うもののように存在しているものが数多く、UMLS の definition を参照しつつ、検討したところ、そのままで使用できるものは数少なかった。また、「日本医学会用語集」には存在しない、MeSH の用語（同義のものも含む）は約 15,000 あった。

D. 考察

シソーラスばかりではなく、辞書に関しては改定年月が長ければ長い程、使用されない用語がそのままの形で残ってしまうので、検索ばかりでなく、投稿時の用語の形態と異なるものがいつまでも存在してしまう。細かな改定作業あるいはメンテナンス作業は必須でなければ、用語集そのものとしての存在価値がなくなってしまうことが考えられる。

E. 研究発表

1. 論文発表
   なし。
2. 学会発表
   なし。

F. 知的財産権の出願・登録状況（予定を含む。）

1. 特許取得
   なし。
2. 実用新案登録
   なし。
3. その他
   なし。

厚生労働科学研究費補助金 （医療技術評価総合研究事業）
分担研究報告書

UMLS と連携する日本語医学用語シソーラスの実用性に関する評価研究
UMLS(Unified Medical Language System)のPubMed(MEDLINE)における検索機能への適応

分担研究者　野添篤毅（愛知淑徳大学）

研究要旨

UMLS(Unified Medical Language System) が PubMed (MEDLINE) 検索で如何に有効に機能しているのかを解析し、主として文献検索で必要な検索機能やシソーラス、検索用語に関する課題、問題点などを明らかにする。シソーラス構築上の課題から日本医学会医学用語辞典改訂に際してのシソーラス化の留意点を明確にすることを図った。

1．PubMed の検索機能と UMLS

PubMedには数多くの様々な検索機能が用意されている。それらは大きく3つに分けられる。まず，Feature barに用意された機能，MEDLINE検索とは異なるPubMed独自の用語自動マッピング機能Automatic Term Mapping，そしてMeSH Browserである。（なお，PubMedの各種の機能の詳細については，NLMがホームページ上で公開しているマニュアル類，PubMed Helpを参照すると良い。）

1.1 Feature bar

PubMedの検索は，Query boxに質問語句を入力することによって実行されるが，その他の検索オプションがFeature barに次の4種の機能として用意されている。

- Limits
- Preview/Index
- History
- Clipboard

Limits

ここでは，MEDLINEの索引作成で用いられるタグ（Check tag）の，研究対象となった年令層，性別，ヒト／動物についての限定ができる。その他，論文の言語，発表年月，入力年月も限定可能である。EBMに関連する出版タイプ（Publication Type），臨床試験Clinical Trial，メタアナリシスMeta-Analysis，診療ガイドラインPractice Guideline，ランダム化比較試験Randomized Controlled Trialなどを指定することができる。また，検索語をMeSH用語と限定することもできる。

Preview/Index

ここには，PreviewとIndexの２つの機能が含まれている。Preview/Indexを用いると次の機能が得られる。

- 文献データを表示する前に検索文献数を見ることができる（Preview をクリック）。
- Query box の検索語句に他の語句を追加することによって検索式を改良することができ，かつ検索文献数を見ることができる（Preview をクリック）。
- Preview/Index を選択した後，All Fields をプルダウンすることによって特定の検索フィールドを指定して検索を実行することができる。これによって，例

えば検索語を MeSH 用語に限定することができる。

- フィールドを MeSH Terms に指定した後，Index ボタンを選ぶことによって，その MeSH 用語とその周辺の語の Index ファイルが文献の生起頻度数とともに示される。これをもとに検索ロジック（AND, OR, NOT）を用いて MeSH 用語による検索式の組み立て，そして検索が実行できる。

History

Historyをクリックすることによって，それまでに実行された検索式と検索文献数が検索順に表示される。これを用いて，CD-ROMでの検索のように検索集合番号とBoolean演算子を用いて，新しい検索式を組み立てて検索が実行できる。

Clipboard

この機能を用いると，いくつかの複数の検索を実行した後に，それぞれの検索を一時的に保管し，後で（1時間以内）ソートなどの処理をして出力することができる。

1.2 Automatic Term Mapping

Query boxに入力されたフィールド指定のない質問語句は，通常PubMedに予め蓄積されている各種の辞書に照合された後，検索ルールによって検索式が自動的に生成される。この際，質問語句から予め決められたStop wordが除かれる。この機能は，これまでのMEDLINE検索とは全く異なるPubMed検索の一番の特長である。質問語句の辞書への照合の順番は次のとおりである。

1. MeSH Translation Table
2. Journals Translation Table
3. Phrase List
4. Author Index

まず，質問語句はMeSH Translation Table（MeSH変換テーブル）と照合される。このテーブルには，MeSH用語，See参照，副標目（subheadings），医学分野の統合用語ファイルUnified Medical Language System（UMLS）（同義語を含む），化学物質名ファイルに登録されている語などを含んでいる。例えば，“vitamin h”と入力すると，変換テーブルからは，“Biotin”[MeSH Terms]と“vitamin h”[Text Word]が示される。そして，検索式としては，

“Biotin”[MeSH Terms] OR “vitamin h”[Text Word]

が自動的に提示される。

Journal Translation Tableでは，入力された雑誌名（正式誌名，略誌名など）がPubMedの略誌名に変換され，検索に用いられる。MeSHあるいは雑誌名変換テーブルと照合できない場合は，次のPhrase Listを探しに行く。Phrase Listには，これまでGrateful Medの検索実験などで収集された数十万の語の組合せリストが蓄積されている。語句の源はMeSH，UMLSであることはいうまでもない。最後に，著者ファイルAuthor Indexと照合される。

検索式は，質問語句と各種辞書から照合された検索語は，同一テーブル，例えばMeSH変換テーブルで選択された用語が同一のMeSHカテゴリーであれば，論理和ORで結ばれ，その他のファイルから選ばれた語とは論理積ANDで結ばれる。例えば，

“head lice shampoo”（アタマジラミ用のシャンプー）と入力すれば，
（“Pedicles [MeSH Terms] OR head lice [Text Word] AND shampoo [All fields]）
の検索式が自動的に組み立てられる。

「インターフェロンによるC型肝炎の薬物療法」に関する検索を行った場合，質問語句からStopword Ignoredに示されるようにofとusingが

取り除かれ，drug therapy，hepatitis C，interferonに分解され，辞書と照合される。Detailsをクリックすることによって PubMed Queryのフレーム中に検索式が表示され，検索文献数がResultsに，語句の変換結果が，Translationに示される。この自動マッピングによる検索では，Query boxへの語句の入力の仕方（語の組み合せ）によって検索結果に差異がでる場合があるので，常にどのような検索語，検索式が使われているのかをDetailsによって確認する必要がある。

2. MEDLINE の索引手法と MeSH Browser

PubMedの基となるMEDLINEデータベースでは，各々の文献は統制用語集Medical Subject Headings（MeSH）によって索引語が与えられる。シソーラスMeSHには，アルファベット順リスト（Annotated Alphabetic List）のほかのカテゴリー別階層構造リストTree Structuresが用意されている。MeSHの登録された約22,000のディスクリプタは，AからNのカテゴリーに分類され，各々のカテゴリーの中でサブカテゴリー細分され，階層構造リストに排列される。カテゴリーは，カテゴリーAは解剖，Bは生物，Cは疾患，Dは化学物質と薬物，Eは診断・治療の技術，Fは精神医学などとなっている。Tree Structuresでは，例えば，眼（Eye）については，

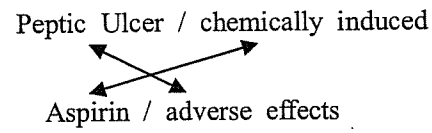
Eye（眼）
  Eyebrows（眉毛）
    Eyelids（眼瞼）
      Eyelashes（睫毛）

のようなMeSH用語の階層構造が組み立てられている。

PubMedでMeSH用語を使って検索すると，自動的にこの階層構造の下位にある用語をもすべて検索の対象としていく。上の場合EyeをMeSH用語として，検索すると，階層構造の下位にある3つの語で索引された文献をも検索することになる。この機能をexplode（拡張）と呼んでいる。

MEDLINEの索引作業では，論文内容をMeSH用語で表す際に，研究の主たるテーマMajor Topicとそれ以外のMinor Topicに重みをつけて索引語を与えていく。各文献につけられた索引語のうちMeSH Major Topicにはアスタリスク（＊）が付けられている。

またMEDLINE索引作成のもう一つの特長は副標目（subheading）の採用である。副標目は，各文献に付与された主標目（main heading）であるMeSH用語が文献中でどのような研究アスペクトで議論されているかを示すタグである。したがって，副標目は必ずMeSH用語と組み合せて用いられ，単独では使用されない。例えば，「アスピリン服用による副作用で起こった消化性潰瘍」という文献については，次のように索引語が与えられる。



すなわち，化学物質Aspirin（MeSH用語）の副作用adverse effects（副標目）によって，消化性潰瘍Peptic Ulcer（MeSH用語）が化学的に誘発chemically induced（副標目）されたという関係が副標目を用いることによって明確に表される。

PubMedではこのような索引手法を基礎にしたこれまでのMEDLINE検索法と同一の検索機能がサイド・バーにあるMeSH Browserを開くことによって提供されている。MeSH Browserの機能は次の通りである。

- MeSH 用語を階層構造で表示する。
- 検索のために MeSH 用語を指定する。
- MeSH 用語を Major Topic のみに限定する。
- MeSH 用語に適切な副標目をつける。

検索語をMeSH用語として指定することは，Feature BarのLimitsでも可能であるが，MeSH Br

owserでMeSH語を探すことによって、入力した語Peptic Ulcerを中心とした階層構造を，上位，下位のMeSH用語を含めて見ることができる。これによって，検索語を広げたり（explode），あるいは狭めることが可能となる。ここには，対象となったMeSH語の定義も示される。次にDetailed displayをクリックすることによって，MeSH語と組み合せ可能な副標目が一覧される。ここでMeSH語に検索質問に対応した副標目を１つあるいは複数選択する。副標目は現在，82種が用意されているが，副標目にもグループと階層関係があり，例えば治療（therapy）に関連する副標目は次のものがあるので，検索の際には，これを考慮するとよい（副標目の階層関係については，各種マニュアル，あるいはPubMed HelpのSubheadings & Families of Subheading Explosionsの項を参照）。

therapy 治
療
diet therapy
食事療法
drug therapy
薬物療法
nursing
看護
prevention & control
予防と抑制
radiotherapy 放
射線療法
rehabilitation リ
ハビリテーション
surgery
外科手術
transplantation
移植

MeSH Browserで選択したMeSH用語は，下位語がある場合には，通常explodeで検索されるため，explodeをしないでその該当MeSH用語のみで検索を実行する場合には（図3）の最下段の“Do not explode this term ”を選択する。また，Major TopicのMeSH用語のみを指定するには，その上段のカラムを選ぶ。このブラウザーによって質問に適応したMeSH用語を検索語として選び，論理演算子でこれらを組み合せて，検索式を組み立て，MEDLINE索引規則に基づいた確かな検索を行うことができる。

G．研究発表

1. 論文発表
なし。
2. 学会発表
なし。

H．知的財産権の出願・登録状況（予定を含む。）

1. 特許取得
なし。
2. 実用新案登録
なし。
3. その他
なし。

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
小野木雄三	UMLSとMedlineを利用した日本語医学用語への意味属性付与	医療情報学	25 Suppl	933-936	2005
鈴木博道他	セマンティックWebエンジンを用いたMedDRA/Jのオントロジ化とその応用方法の研究	医療情報学	25 Suppl	1317-1319	2005

UMLSとMedlineを利用した日本語医学用語への意味属性付与

○小野木 雄三[1)]
東京大学大学院医学系研究科クリニカルバイオインフォマティクス研究ユニット[1)]

Finding semantic information for terms in Japanese medical dictionary using UMLS and Medline

○ONOGI YUZO[1)]
Department of Clinical Bioinformatics, Graduate School of Medicine, the University of Tokyo, Tokyo, Japan[1)]

Abstract: To analyze clinical documents written in natural language or to process some intelligent tasks in computers, it requires semantic information for each medical term for targeted domain. Then we tried to extract relevant semantic categories for each Japanese medical term in Japanese medical dictionary published by the Japanese association of Medical Sciences using Medline titles and abstracts. For Each English term in the dictionary, we calculated weight matrix by TF-IDF method from frequencies of terms and MeSH indexes appeared in Medline articles from year 2000 to 2004. About 50,000 terms in the dictionary (which has 80,000 terms) were found and could successfully assign semantic categories. To evaluate this relevancy, we compared relevant MeSH terms, and their assigned semantic categories with dictionary terms which are MeSH terms by themselves, and got accuracy of about 76%.

Keywords: UMLS , MEDLINE , Japanese Medical Term , Semantic Information , TFIDF method

1. 背景

語彙あるいは概念に付随する意味情報は、自然言語処理、テキストマイニング、知的情報処理などに有用である。例えば胆嚢、胆嚢腺筋症、胆石などの語彙は、それぞれ解剖学的部位名、疾患名、疾患名や所見として使われること、などが解っていれば、解析が有利である。しかし日本語医学用語に対する意味情報付与は、MeSHやICD10など統制用語集へのマッピング[1) 2)]、メルクマニュアルからの知識抽出[3)]、あるいは退院時サマリーに出現する語彙を基にした類似症例検索[4) 5)]などで行われているが、その他の大多数の日本語語彙にはこのような意味的関連付けが存在しない。(医学概念に関しては、MeSHからの概念間関係抽出[6) 7) 8)]やオントロジー構築[9)]などの研究が活発に行われている。)そこで我々は、日本医学会の医学用語辞典(英和)、米国NLMのUMLSおよびMedlineを利用することにより、辞典に収載されている日本語医学用語に意味的な属性を関連付けることを試みた。

2. 目的

医学用語辞典に収載されている日本語医学用語に意味的な関連づけを付すこと。

3. 材料

対象とする日本語医学用語は、医学用語辞典和英(日本医学会)の日本語見出しを利用した。MedlineはNLMより2004年以前のarchiveをXML形式で入手、UMLSはNLMより2005AAを使用した。

4. 方法

医学用語辞典(和英)の日本語見出しから、それに対応する英語見出しを取得する。ここで日本語見出しには複数の英語見出しが対応する場合もあるが、得られる全ての英語見出しを使用した。その英語見出しを小文字に変換したものと、Medline文献(タイトルとアブストラクト)の 2000年以降2004年までに出現する語彙をやはり小文字に変換したものとを比較し、一致するものを対象とした。Medlineでは個々の文献にMeSHタームが索引付けされているため、TFIDF法により日本語見出しとMeSHとの関連を得ることができる。図1にタームと文書に対する重み行列の計算法を示す。ここでiはターム、jは文書、Nは全文書数、niはタームiを含む文書数である。この場合にはタームが英語見出し、文書がMeSHタームに対応する。重み行列が得られれば、各英語の見出し語に対応する日本語見出しと、それに対応するMeSHタームとを、関連の高い順に取得することができる。次にMeSHタームからMeSHのカテゴリー分類を利用して意味的情報を取得、あるいはUMLSを介して他の統制用語の意味的分類やセマンティックネットワークでの意味分類を取得することができる。なお、例えばMeSHやセマンティックネットワークではひとつの概念が複数の階層カテゴリーに分類されている。従ってひとつの見出し語に関連するMeSHタームが得られると、そのMeSHタームには複数の階層カテゴリーが対応することになる。また、UMLSを介して関連付けをMeSHから他の統制用語に変換する際には、対応関係の存在しない概念が多く存在する。その場合には得られる意味的関連は、対応関係が存在しない数に応じて減少する。例えばMeSHとSNOMED-CTではUMLSのCUIが完全に一致する概念が比較的少ないため、対応関係を増やす目的でSNOMED-CTのisa階層構造において上位5階層以内に一致するMeSH概念が存在する場合には同じ概念とするようなマップを作成して利用した。

ここでTFIDF法により文献に出現する英語見出しとMeSHタームとの重み行列を計算する際に、2つの重み行列、すなわち見出し語と文献、文献とMeSHタームの重み行列を得た後に、その積(文献空間における見出し語ベクトルとMeSHベクトルの内積)を計算する方法を取った。以上により、日本語医学用語(見出し語)と意味情報との関連づけを行った。

次に、関連性の評価を行うために、医学用語辞典の英語見出しの中でUMLS、特にMeSHに一致する語彙を取得した。英語見出しにnormalize処理を行い、UMLSのMRXNS_ENG(UMLSの英語語彙をnormalize処理したもの)と比較する。一致したものの中から特にMeSH(同義語も含む)に属する語彙を使うことにより、上記で得た重み行列の中で英語見出しがMeSHに一致するものと、それと関連性の高いMeSHの上位1位から5位までとを比較することによって一致度を評価した。なお、normalize処理でUMLSと一致した英語見出しは、少なくともUMLS中の何らかの統制用語に対応するので、その統制用語の階層構造からその語彙の意味属性を決定することもできる。この方法では意味的関連と言うよりも、表記の完全一致から得られる概念に対応する、より正確な意味(これは統制用語に依存するが)を取得することになる。

5. 結果

医学用語辞典の見出し語数は約8万語彙であった。この中で2000年～2004年(5年分)のMedline文献中に出現した見出し語の数は49384語彙、辞典見出し語の62%をカバーすることができた。重み行列の計算により、これら約5万語の語彙に対して意味的関連付けを付すことができた。この結果の一部を図2に示す。ここで辞書見出し欄は辞書の日本語見出し、MeSH_JPNは関連性が1位のMeSHタームをUMLSのMRCONSOでsourceがMSHJPN(日本語版MeSH)で表示したもの、MeSH分類(1)はMeSHの最上位の分類名称(階層名)を英語で表記したもの、MeSH分類(2)はその1段下位の階層の概念名を英語で表記したものである。なお、試みに1995年以降の10年分で計算を行ったところ、53670語彙(辞典見出し語の67%)が出現していた。

次に、辞典見出し語の中でnormalize処理によってUMLSと一致した語彙数は33487、約3万語であった。統制用語集ごとの内訳は、SNOMEDCTが22020、MeSHが15812などであった。意味的関連付けの精度を評価するために、Medline文献中に出現した英語見出しで同時にMeSHタームであるものを対象として、関連づけられた