別添２

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と連携した日本語医学用語シソーラスの作成に関する研究

平成１４年度　総括・分担研究報告書

主任研究者　脊山　洋右

平成１５（２００３）年４月

別添３

目　　次

別添４，５

厚生労働科学研究費補助金（医療技術評価総合研究事業）
総括研究報告

UMLS と連携した日本語医学用語シソーラスの作成に関する研究

主任研究者　脊山洋右　お茶の水女子大学教授

研究要旨　医学情報流通の標準化には日本語医学用語シソーラスを開発することは必須条件である。米国で開発された Unified Medical Language System(UMLS)と日本医学会が作成した医学用語辞典及び医学中央雑誌刊行会による医学中央雑誌シソーラスその他、既存の用語集などを活用し、比較的短期間で、日本語医学用語シソーラス開発の方法論を確立する。本年度は、UMLS 用語に対する日本語（機械的）対応の確認、非対応語の対応付け、これらをサーバー上で分担し実施するための作業用インターフェイス作成、を行った。

分担研究者
野添篤毅　愛知淑徳大学教授
大江和彦　東京大学医学部附属病院中央医療情報部長（教授）
菊地優子　国際医療福祉大学助教授
佐々木哲明　（財）医療情報システム開発センター普及調査部長
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事業推進室長

A. 研究目的

正規化した英語文字列による機械的マッチング作業により、どの程度まで実現可能であるかの量的な把握が、まず第一の目的である。

機械的マッチング作業のみでは、個々の用語に複数の意味・概念が存在することから UMLS との対応には適正で無いものも存在するはずである。この作業は、人手でやらざるを得ないものであることから、これら点検作業を分担研究者や研究協力者が分担し実施し成果を共有出来るような Web 上のインタフェイスを開発、試用し、ソフトウェアの改良を図って行く。結果として、これまで試みてきている機械的マッチング作業自身の評価、その結果に関する質的な評価、も、合わせて得られる筈である。

B. 研究方法

データベース管理システムには RedHat Linux 上で postgreSQL を、Web サーバーには Apache version 1.3.27 を、両者の連携には php version 4.2.3 を、そして当初の作業データとして UMLS2002 と医学中央雑誌刊行会医学用語シソーラスと

Mesh2001 とを利用した。

日本医学会医学用語辞典の用語による UMLS への機械的マッチングでは 50%弱が、UMLS へマッチングされたが、このマッチングされたものの妥当性をサーバー上で分担して点検し、問題点を抽出し評価した。

また同時に、UMLS に対して自国語の翻訳版を提供している各国の情況についても、その現状把握を、文献並びに現地調査によって行った。

（倫理面への配慮）

本研究では医学用語を対象とするのみであり、倫理面での配慮は必要ない。

C. 研究結果

既存の日本語医学用語集などに収容されている用語について、英語表記を介した機械的なマッチングにより UMLS の概念にリンク可能なものは 35 %程度であり、英語の単語単位でのマッチングその他、補足的な手段を講じて上でも、45 %以下に留まることが判明した。スペルチェックをかけることにより、わずかながら、マッチング率は向上した。

以上の作業の結果をまとめたファイルを作成し、サーバー上で班員が参照できるようなシステムを構築した。1 つの概念(コンセプト)に対し、機械的マッチング作業で一致した用語やその表示形が、日本語訳などと同時にその出典も含めて参照できる。2 年目以降、追加する用語集からの入力を済ませた段階で、班員および協力者が分担し、英語コンセプト、各種用語集の英語見出語と日本語訳、などを対照させつつ、機械的マッチングの結果を点検評価することが可能となる。

UMLS に関する文献調査の結果などから、現段階での UMLS の活用は基礎的研究から診療記録への活用まで、幅広く実施されていることが明らかとなった。全 316 件の UMLS 研究論文中で、言語や用語に注目した基礎的な研究は 74 件わずか 23 %程度であり、応用的な研究が 230 件(73 %)を占めており、UMLS が単なる基礎的実験研究レベルでは無いことを明らかにしている。応用例の中でも、電子カルテ・診療記録などへの応用が文献検索への応用をはるか 3 倍以上、上回っていることも注目に値する。UMLS を提供している NLM との交渉により、本研究の UMLS 研究プロジェクトに対する協力と同時に、NLM からの本研究に対する協力もとりつけることが出来た。

D. 考察

対象とした医学用語集などの英語文字列など表記統一やスペルチェックと訂正、用語集や UMLS に付与されている各種コードを活用した準機械的マッチング、などの手段を講ずることで、全体の 50%程度の機械的マッチングが期待できることが実証された。大量の用語の迅速な概念的リンク実現には、機械的手段の存在が重要であることは明白であり、当初の調査で対象とした既存医学用語集などと

その英訳語の整備を行った上、再度機械的なマッチングを実施することが必要とされている。

同時に、機械的マッチングのみでは不十分なことから、一定の領域を限定した上での手作業によるマッチングを試行し、その効率向上を図ることも、必須である。

E. 結論

今後の研究計画として以下を予定している。

(1) 日英対訳の医学辞書・用語集などから収容されている用語を１つのファイルにまとめる。電子化されていない用語集についても早急にこれに加える。データは、英語／日本語訳のペアに出典を示すキー、が中核となるが、現状では日本語には複数の見出語が対応するものも多い。この整合性のチェックは後の課題とする。

(2) ユニークな英語見出語を取り出し、スペルチェック、正規化処理を加えた上で、UMLS とのマッチングを実施する。この結果、約 50 %弱の語がマッチングされることとなる。日本語見出語でコンセプトに対応しない語が 50,000 語、対応する語が 50,000 語と予想される。

(3) 日本語見出語でコンセプトに対応しない語は、その医学用語英名を NLM に送付し、UMLS への追加登録を含めた検討と理由の解明を依頼する。何らかの理由が明らかとなり、パターンがある場合は、これまでの作業をやり直すこともあり得る。

(4) 日本語見出語でコンセプトに対応した語についても、単なる機械的マッチングの成果でしかないことから、人手によるチェック作業を行う。この過程で、ファイルの整備と同時に、典拠となった用語集などの点検も、実質的に行われることとなる。この段階では、原則的に削除は行わないものとする。

(5) 以上(2)から(4)のステップを、全分野で行うことは困難が予想されることから、Semantic Type: Disease or Syndrome の一部に限って行う、などのサンプル・データで実施する。

(6) 対応付けのチェック作業を終えたものに対して、テストを行い、評価する。

F. 健康危機管理情報

なし

G.研究発表

特になし

H.知的財産権の出願・登録状況

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）
総括・分担研究報告

UMLSと連携した日本語医学用語シソーラスの作成に関する研究
UMLSと日本語医学用語との対応付け

主任研究者　脊山洋右　お茶の水女子大学教授
分担研究者　大江和彦　東京大学医学部附属病院
研究協力者　波多野賢二　東京大学医学部附属病院

研究要旨　これまで日本医学会医学用語辞典を使い、英語文字列を normalize することによる機械的な一致を行ってきたが、これに加えて医学中央雑誌刊行会の「医学用語シソーラス第5版」およびMEDIC-DCの「ICD10対応電子カルテ用標準病名マスター」を利用して対応の拡大を図ることができることが考えられる。ここではそれを実施した過程と結果について報告する。

分担研究者
野添篤毅　愛知淑徳大学教授
大江和彦　東京大学医学部附属病院中央医療情報部長（教授）
菊地優子　国際医療福祉大学助教授
佐々木哲明　（財）医療情報システム開発センター普及調査部長
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事業推進室長

A. 研究目的

UMLS (Unified Medial Language System) Metathesaurusと日本語医学用語との対応付けを行うことが当研究班の目的である。

B. 研究方法

日本語医学用語のリソースと UMLS Metathesaurusにはそれぞれ以下のものを使用した。個々のリソースの内容をデータベースに登録し、作業を行った。データベース管理システムには、RedHat Linux version 7.3 上の postgreSQL 7.2.3 を使用した。

- 米国 NLM が管理・配布する UMLS2002 年 AC バージョン(秋版)
  ここには 870,853 個の概念が収容されている。これらの概念に対応して同義語と異表記の語彙が合計 2,083,103 個存在する。これには英語だけでなく 15 カ国の医学用語が含まれており、英語の語彙だけに絞ると 1,753,789 個である。

- 医学中央雑誌刊行会の医学用語シソーラス第5版　(医学用語シソーラス)
  このシソーラスは統制語(概念)として55,907個を有し、それらに対応して同義語が28,702個、類義語が95,930個、関連語が124,045個、合計で304585語の語彙を収載する。この統制語のうち、MeSH2001年版MeSH Heading英語文字列に対応しているものは19,174個であり、同義語・類義語・関連語の数はそれぞれ26,844個、95,059個、32,505個、合計で173,583個になる。ここで収載されている用語は、日本で刊行されている医学関連文献で使用されているものを基にしているが、日本語の文献には英語文字列も常用されているため、このシソーラスにも英語文字列が含まれている。そこで日本語の語彙だけに絞ると、その用語数は74,706個となる。これを表1に示す。ここで、MeSHに対応していないものはすべて統制語を持たず、同義語～関連語だけで構成されている。これはフリーキーワードと呼ばれており、概念としてはひとまとまりになっているが統制語が割り付けられていない状態の語彙であり、いずれ統制語が割り当てられ、MeSHにも対応付けられる予定のものである。
- ICD10対応電子カルテ用標準病名マスター　(標準病名マスター)
  財団法人医療情報システム開発センター(MEDIC-DC)が配布するもので、ここに収載されるすべての日本語病名はICD10の3桁および4桁コードに対応している。病名として19,776個の語彙を有し、それらに対応するICD10コードの種類は6,674個である。この中には、ひとつの病名に対して複数のICD10コードが割り当てられているものもあるが、ここではすべてひとつの病名にひとつのICD10コードが対応するように展開して利用した。
- 日本医学会医学用語集 英和辞典(2001年版)　(医学用語辞典)
  見出し語の英語語彙数は83,791語であり、これに複数の日本語が対応する。見出し語および日本語には各種の記号（上付き文字、下付き文字、言い換え、付加）が存在し、さらに英語語彙の中にギリシア文字など全角の日本語文字が含まれているため、前処置が必要であった。

これらのリソースをUMLSに対応付ける際の方針を図1に示す。医学用語シソーラスはMeSHに対応しているので、UMLSのMRCON（語彙テーブル）とMRSO（出典テーブル）からMeSH Headingの語彙を探し、これと医学用語シソーラスの統制用語の英語と一致するものを結合することによって、各統制用語とその同義語・類義語・関連語をUMLSの概念に対応付けることができる。同様に、標準病名マスターはICD10に対応しているので、その3

桁・4 桁コードをもとに MRSO と MRCON から対応する UMLS 概念を結合することができる。医学用語集については、従来の当研究班の成果、すなわち医学用語集の見出し語を normalize した英語文字列と MRXNS（UMLS Metathesaurus

に含まれるすべての語彙を含む MRCON の文字列を normalize したもの）を単純に比較して一致するものを探すことにより、対応付けを行うことができる。この様子を図 1 に示す。

C. 研究結果

日本語リソースごとに、UMLS 概念との対応付けを以下のように行った。

- 医学用語シソーラス

  対象語彙は「MeSH に対応している統制語・同義語・類義語」である 141077 語とした。表 1 の太字で示した部分にあたる。関連語まで含むと UMLS 概念との対応があいまいになってしまうと考えたため、関連語を除外した。また MeSH に対応していない語彙は、UMLS との対応を取ることができないために除外している。

  医学用語シソーラスの各統制用語には 2001 年版 MeSH heading の英語文字列が対応している。そこで UMLS 2002 年秋版の MRSO から MeSH をソースとする文字列を抜き出し（20743 個）、それと医学用語シソーラスの英語文字列との一致を行うことによって UMLS の CUI に対応させた。これにより、統制用語 19174 個のうち 19128 個で CUI と対応付けることができた。対応しなかった 46 個は MeSH 2001 年版と 2002 年版との違いによるものと考えられる。また 2002 年版 MeSH には 20743 個の概念が存在し、医学用語シソーラスの概念数(19174 個)は 1600 個ほど少ないことになるが、これは地理情報を表す Z 軸(368 個)のほか、日本語に対応するものが存在しない概念による影響と考えられる。

- 標準病名マスター

  複数 ICD コードを持つものを展開した上で、ICD10 コードとして 3 桁または 4 桁のものが存在する病名を対象としたため、対象数は 18964 個となった。標準病名マスターには、病名に対応する ICD10 コードが記載されているため、これをもとに UMLS の概念と対応付けることができる。そこで UMLS 2002 年秋版の MRSO から ICD10 をソースとするコードと概念を抽出し(12027 個、概念数は 11357 個)、標準病名マスターの ICD10 コードと一致させることによって日本語病名に対応する UMLS の CUI に対応させた。その結果、対応する CUI を見出せなかったものが 2 個、それ以外の 18962 個の日本語病名は UMLS の概念(6617 個)に対応付けることができた。なお、上記 2 個は「両側性硬口蓋裂　Q350」と「両側性軟口蓋裂　Q352」で、いずれも対応

するICDコードがUMLS側に存在せず、それぞれQ351とQ353が対応していた。

医学用語集

最初に、英語文字列あるいは日本語文字列の中に含まれている記号類を削除または展開した。表2にそれらの例を示す。もとの見出し語数は83,791個であったが、この展開操作によって英語語彙数は86,142個に、日本語語彙数は102,619個になった。また展開によって生じた語彙の重複を除くと、英語語彙数は80,327個で日本語語彙数は87,759個であった。

次に英語文字列をUMLSのlvg (Lexical Variant Generation)ツール中のnormを使って標準形に変換した。このnormは文字列をすべて小文字に変換し、英語のストップワードや格変化・語形変化を取り除き、単語をアルファベット順に並べ替える。そのため、ひとつの英語文字列を処理すると複数の結果が出力される。例えば"feeling of drunkenness"の場合には、" drunkenness feel"と" drunkenness feeling"が出力される。その結果、変換後の英語標準化文字列数は95,979個、重複を除いて88,467個となった。

次にUMLSの統制用語をnormで処理した結果であるMRXNSとの比較を行い、標準化した英語文字列が一致したものに対してCUIを対応させた。119,236個の対応が得られたが、重複を除くとCUIとしては29,898個、もとの英語文字列では33,333個、英語見出し数では36,118個が対応した。

最後に元の辞書の見出し語の英日対応をもとに、CUIと日本語語彙とを対応させた。単なるjoinでは146,017個の対応が得られたが、見出し数で見ると36,118個、日本語語彙では39,283語であった。

一応、もとの辞書の見出し語のうち約45%がUMLSと対応付けされたことになるが、normの処理も含めて機械的な一致による誤対応が発生している可能性に留意する必要がある。

以上の作業の結果を表3に示す。これは単に3種類の日本語リソースごとにUMLSとの対応作業を行った結果なので、同一の日本語語彙がリソース間で異なるUMLS概念に対応して、全体として矛盾を生じている可能性がある。そこで3種類のリソースに現れた日本語語彙をすべてひとつにまとめ、個々の語彙がどのUMLS概念に対応しているのかを調べる必要がある。同時にこのひとつにまとめた日本語語彙集は、UMLSに対応する複数の日本語語彙がリストされるシソーラスであるため、なるべく多くの日本語語彙を集めておいた方がシソーラスとしては有用であると考えられる。そこで、医学用語シソーラスで使用していなかった語彙（MeSH概念に対応していないが同一の概念として分類されている語彙群および関連語）を加えて大きな日本語語彙テーブルを構築した。これにより、MeSH経由ではUMLSに対応していない日本語語彙が、標準病名マスターや医学用語集経由でUMLSに対応する可能性があり、それによって（日本語語彙の機械的一致として）より多くの日本語語彙をUMLSにマップすることが可能と考えられる。

また、ひとつの日本語語彙がUMLSの複数の概念と対応してしまう場合を除くために、これまで一致した語彙の中から複数のUMLS概念に対応しているものをリソース

ごとに調べて取り除いた。特に医学用語集では英語すなわち UMLS 概念と日本語語彙とは単一の見出しとして対応付けられているため、UMLS と対応の付いた見出しに含まれる日本語語彙をマップした。これにより、UMLS と 1 対 1 で対応の付いた見出し数は 32,665 個となり、それに対応する日本語語彙は 40,181 個、対応する UMLS 概念の数は 25,228 個となった。表 3 に比べて対応する UMLS 概念数が減っているのは 1 対 1 対応のものだけに絞ったからであり、それにもかかわらず対応の付いた日本語語彙数が増えているのは見出し語に含まれる日本語語彙すべてに拡張したからである。

以上により 3 つの日本語リソースに含まれる日本語語彙をすべてあわせた 383,514 個に対して、表 4 を作成した。個々の語彙が UMLS 概念と 1 対 1 に対応している場合だけを数えて'one'と表記し、対応する語彙数を(括弧内は UMLS 概念数)を記載した。'ー'と記載されている部分は、対応する UMLS 概念が全く存在しないことを示す。また複数箇所に'one'と記載されている場合は異なるリソース間で同一の UMLS 概念に対応していることを表す。これにより、単一のリソース内および複数のリソース間で、異なる UMLS 概念に対応してしまう語彙を除くことにより、確実に UMLS 概念と対応する日本語語彙を抽出したことになる。なお、複数の UMLS 概念に対応した日本語語彙数は 2,344 個であり、意外と少ないのであるが、これに対しても対応している UMLS 概念に上位・下位関係が(MRREL などを使って)認められるならば、どちらかを代表させて日本語語彙に対応付ける方法が考えられる。

これをまとめなおしたものが表 5 および表 6 である。個々のリソース内部で UMLS と一致させた場合に比べて、複数の UMLS 概念に一致してしまったものが除かれているため、対応率は悪くなっている。表 6 では日本語語彙として、すべて英数字で構成されるものを除いた場合と含んだ場合を示す。

こうして作成した日本語医学用語と UMLS 概念およびそれに属する英語医学用語との対応表は、作業班のホームページに載せた。日本語の部分文字列を入力することにより、UMLS と対応している日本語医学用語を検索し、同時にそれに対応する UMLS 概念、英語表記、それぞれの表記の出典を提示することができた。

D. 考察

今回の作業として、3 種類の日本語リソースに収載された日本語語彙を、重複を除いて並べて UMLS 概念との一致を図ることにより、異なるリソース間での日本語文字列の一致による UMLS 対応の向上を図った点に特徴がある。しかし実際にはひとつの日本語語彙が複数の UMLS 概念に対応してしまうものを除いたために、結果として一致率の向上ははっきりしなかった。複数の UMLS に対応した語彙の処理方針としては、結果にも述べた通り、UMLS の MRREL を使って概念の同義関係や上位・下位概念を得ることにより、ひとつの UMLS 概念に対応させることは可能であり、今後この作業を行うことにより、この方針の正確な評価が可能になると考えられる。ただし現在のところ、複数の UMLS 概念に対応している語彙数がそれほど多くないこ

とを考えると、あまり有効ではないと言えよう。むしろ UMLS に対応させることのできなかった語彙が依然として多いこと、これを減少させることこそが課題である。ただし、医学用語集には形容詞など、対応するはずのない語彙も多く含まれているため、逆にこのような「対応不能語彙」を特定して除く、という方針の方が有用であるかもしれない。

少なくとも今回の対応作業で約 20 万件の日本語語彙が UMLS 概念と対応したことになり、この点は評価できよう。今後はこれを使うことによって、日本語文字列を入力して MEDLINE 検索を行うなどのアプリケーションを容易に開発することができる。ただしここで注意すべきは、UMLS 概念に対応しているとは言っても、今回の作業で得られた日本語語彙→UMLS 概念→英語文字列から直接に MEDLINE 検索を行うことができない可能性があることである。何故なら MEDLINE の文献検索は MeSH でインデックスされているからであり、今回の対応表で得られた英語文字列が MeSH に属するものでなければ有効ではない可能性があるからである。つまり正確に MEDLINE 検索を行うためには、得られた UMLS 概念に最も近い MeSH 概念を返す様にするべきである。一般的には、要求されるアプリケーションの目的に応じて最適な出典(vocabulary source)の用語を返すようにする必要がある。

E. 結論

医学用語シソーラス、標準病名マスター、医学用語集に収載されている日本語医学用語をすべて集め、その中から UMLS 概念と 1 対 1 に対応する用語を見出した。

F. 健康危機管理情報

なし

G.研究発表

特になし

H.知的財産権の出願・登録状況

なし

表1．　医学用語シソーラスに収載される語彙の内訳。今回の作業に使用したのは、太字で示されている部分で、合計 141077 語である。

	統制語	同義語	類義語	関連語	合計
MeSH 対応あり	**19,174**	**26,844**	**95,059**	32,505	173,583
MeSH 対応なし	36,733	1,858	871	91,540	131,002
合計	55,907	28,702	95,930	124,045	304,585

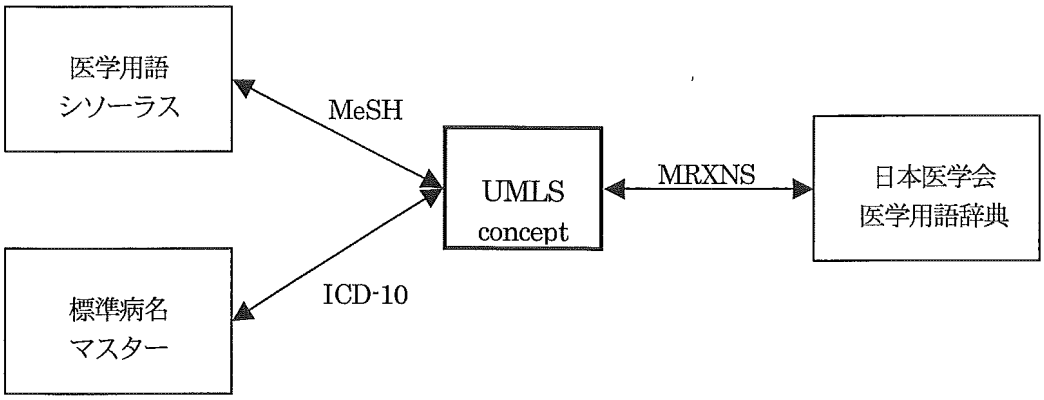

図 1.　医学中央雑誌刊行会の医学用語シソーラス第 5 版の日本語は MeSH を介して、MEDIS-DC の ICD10 対応電子カルテ用標準病名マスターの日本語は ICD-10 を介することにより、比較的容易に UMLS 概念と対応を取ることができる。日本医学会医学用語集英和辞典の日本語は、これまでと同様に normalize した英語文字列を MRXNS テーブルと比較することによって UMLS 概念と対応付けることができる。

表2．　日本医学会医学用語辞典英和辞書の見出し語に存在する記号を削除または置換する際に行った展開方法の一覧。これは英語見出しだけでなく、日本語欄の文字列にも適用した。

記号	展開方法	例	展開後
（）代替語	2 つの entry を生成	Anti-leprosy drug (agent)	Anti-leprosy drug Anti-leprosy agent
〔〕付加可能	2 つの entry を生成	abdominal 〔wall〕 reflex	abdominal reflex abdominal wall reflex
全角文字	半角に置換もしくは削除	α、β	alpha, beta
☆･･･★	削除	☆Acacia★	Acacia
▼下付き文字	削除		
▲上付き文字	削除		
△上付き文字	削除		
外字定義	既存文字で代用	【外 6FEB】	濫

表3． 各日本語リソースの日本語医学用語の語彙数、そのうちUMLS概念と対応の付いた日本語の語彙数、および対応するUMLS概念の数を示す。医学用語シソーラスには、英語文字列も日本語として登録されている。こうした英語文字列も日本語とみなした場合の数を括弧内に示す。

	医学用語シソーラス	標準病名マスター	医学用語集
日本語語彙数	75,630 (141,077)	18,964	87,759
UMLSと対応の付いた語彙数	75,455 (140,721)	18,962	39,283
対応するUMLS概念の数	18,920 (19,133)	6,617	29,898

表4． 日本語医学用語とUMLSとの対応が日本語リソース間で重複する数を示す。ひとつの日本語語彙に対してただひとつのUMLS概念が対応した場合を'one'と記述する。

医学シソーラス	標準病名マスター	医学用語集	語彙数	概念数
one	-	-	157,953	19,123
-	one	-	13,035	5,074
-	-	one	20,245	14,753
one	one	-	379	346
-	one	one	368	358
one	-	one	8,605	6,680
one	one	one	604	574
合計（日本語語彙とUMLSが1対1対応しているもの）			201,189	35,705
1つの日本語語彙に複数のUMLSが対応			2,344	0
対応するUMLS概念なし			174,111	0

表5．UMLS概念と1対1対応した日本語語彙数、各リソースに収載される語彙数、およびそれぞれの対応割合を示す。個々のリソースごとにUMLSと対応させた場合に比べて対応率が低下しているのは、複数のUMLS概念に対応したものが除かれているためである。

	医学シソーラス	標準病名マスター	医学用語集	合計
UMLSと1対1対応した語彙数	167,541	14,386	29,822	201,189
各リソースに含まれる語彙数	304,585	19,776	87,759	383,514
UMLSとの対応率	55%	73%	34%	53%

表6．　収集した日本語語彙数、そのうちUMLS概念と対応付けることのできた日本語語彙数、およびそれに対応するUMLS概念の数を、日本語語彙としてすべて英数字で構成されるものを除いた場合、含んだ場合について示す。

	英語文字列を除く	英語文字列も含む
日本語語彙数	229,163	383,514
UMLSと対応の付いた語彙数	122,025	201,189
対応するUMLS概念の数	35,370	35,705

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

UMLS と連携した日本語医学用語シソーラスの作成に関する研究
UMLS MeSH 用語に対する日本語対応の確認と、作業用インターフェイスの作成

主任研究者　脊山洋右　お茶の水女子大学教授
分担研究者　大江和彦　東京大学医学部附属病院
研究協力者　小野木雄三　東京大学医学部附属病院

研究要旨　当研究班ではこれまで日本医学会医学用語集（英和辞典）と UMLS (Unified Medical Language System)との対応付けを、normalize した英語文字列を機械的に比較することにより行ってきた。しかし辞書の見出し語には複数の意味が存在するため、統制用語集である UMLS との対応には適正でないものが生じる可能性がある。この不整合を排除する作業は機械的に行うことはできず、人間が行わなくてはならない。しかし医学中央雑誌刊行会の医学用語シソーラス第 5 版は、それ自体が統制用語集であり、しかも MeSH との対応がついている。これを利用すれば日本語医学用語と UMLS Metathesaurus とを整合性を保って対応付けすることができる。そこで MeSH を介して、医学用語シソーラスに収載されている日本語医学用語と UMLS 概念との対応付けを行い、これが正しく対応しているか否かを確認することとした。

分担研究者
野添篤毅　愛知淑徳大学教授
大江和彦　東京大学医学部附属病院中央医療情報部長（教授）
菊地優子　国際医療福祉大学助教授
佐々木哲明　（財）医療情報システム開発センター普及調査部長
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事業推進室長

A. 研究目的

医学用語シソーラス第 5 版の日本語と UMLS 概念とが正しく対応していることを確認し、対応していない場合には日本語医学用語を修正することが目的である。ここで研究班のメンバー全員でこの作業を行うために、インターネットを介して確認作業を簡便に行うことのできるインターフェイスを構築することが必要である。

B. 研究方法

1. 材料と方法

日本語医学用語のリソースと UMLS Metathesaurus にはそれぞれ以下のものを使用した。

- 米国 NLM が管理・配布する UMLS2002 年 AC バージョン(秋版)
- 医学中央雑誌刊行会の医学用語シソーラス第 5 版(MeSH 2001 年版と対応)

この 2 つの用語集を登録するデータベース管理システムには、RedHat Linux 上の postgreSQL を使用し、Web サーバーには Apache version 1.3.27 を、Web サーバーとデータベース管理システムとの連携には php version 4.2.3 を使用した。

1.1. 医学用語シソーラス

医学用語シソーラスのデータベースは、統制語テーブル、辞書テーブル、典拠リストテーブル、典拠テーブルの 4 つから成る。この構造を図 1 に示す。

統制語テーブルでは tosei_c がユニークで、医学概念ごとに 1 レコードが存在し、この概念に対応する複数の医学用語が辞書テーブルに収められている。また統制語テーブルの nlm_y は MeSH Heading(英語文字列) に対応するので、この英語文字列を介して UMLS の概念を表す CUI (Concept Unique Identifier)に対応付けることができる。(この様子を破線で示す)。

辞書テーブルにはひとつの概念に対応するさまざまな医学用語が登録されているが、そのレベルが dogi_no で規定される。dogi_no=0 はその概念のリードターム、dogi_no=1 は同義語、dogi_no=2 は類義語、dogi_no=3 は関連語となっている。Seq_no は dogi_no が同じ群の用語を 0 から順に指定するものなので、tosei_c、dogi_no、seq_no が定まれば医学用語がひとつ定まることになる。

典拠リストテーブルは、個々の医学用語がどの典拠資料に存在していたかを示すものであり、ひとつの用語に対して id_no で指定された数だけの典拠が存在する。個々の典拠資料は tenkyo_c によって典拠テーブルの典拠資料名称に対応付けられる。

1.2. UMLS

UMLS の中でこの作業に使用するテーブルの構造を図 2 に示す。MRDEF、MRCON、MRSO の 3 つのテーブルである。

MRDEF は CUI に対応する概念の定義を収容する。SAB には出典(統制用語のソース) が記載されており、同じ概念に対して複数の出典がある場合には出典ごとに別の定義が記載されている。

MRCON は UMLS Metathesaurus の主要テーブルであり、ひとつの CUI すなわち概念に対応するすべての用語を STR に収容する。ここで LUI、SUI はそれぞれ Lexical Unique Identifier、String Unique Identifier を意味する。複数形の違いなどのように 1 文字が異なるだけでも異なる SUI を持ち、表記が似たものは同じ LUI を持つ。いずれも同じ概念に属するので同一の CUI を持つ。このことを説明する資料として UMLS からの引用を表 1 に示す。LAT には言語が入る。2002AC には英語、フランス語など 15 種類の言語が含まれるが、STR は 7bit ASCII で表記されている。当作業では英語のみを対象とした。TS は Term Status で、Preferred Name か Synonym かを示す。STT は String Type で、Preferred form of term か Variation かを示し、

Variationの場合には大文字・小文字の違い、順序の違い、単数形・複数形の違い、などを示す記号が収容される。

MRSOは各語彙の出典を示すテーブルである。CUI、LUI、SUI ごとに定まる表記がどの統制用語からのものであるかを示す。ここでSABは出典を、TTYは各統制用語でのタイプを示し、CODEは各統制用語での識別子を収容する。ここではMeSHを対象としているので、SAB="MSH2002_06_01"が対象となる。またCODEにはSABがMeSHの場合にはMeSH Tree Numberが入る。例えばMeSH Heading が"abdominal neoplasms"にはC04.588.033というMeSH Tree Numberが対応する。ただしMeSHではひとつのMeSH Heading あるいは概念に対して複数のTree Numberが存在することに注意する必要がある。例えば" Burkitt Lymphoma" に対してC02 (Virus Diseases)、C04 (Neoplasms)、C20 (Immunologic Diseases)などの軸に属する13個のTree Numberが存在する。

1.3. 医学用語シソーラスとUMLSとの結合

さきに述べたように、医学用語シソーラス第5版のnlm_yは2001年版MeSH Headingに対応する。この文字列は、MRCONのSTR中に存在するーただしそのCUI・SUIはMRSOで出典がMeSHであるものに限るーはずである。そこでnlm_yとSTRが一致するものを結合すれば、医学用語シソーラスのtosei_cとMRCONのCUI・SUIが一意に対応する。あとはtosei_cに対応する日本語の医学用語を同義語レベルまで呈示（類義語以下は参考情報として呈示）して日本語側の情報とし、CUIに対応するMRDEFの定義とMRCONのSTRすべてを呈示して英語側の情報とし、この両者を比較することで確認作業を行うことができる。

問題はUMLSが2002ACであり医学用語シソーラスと版が異なることである。改版によってSTR文字列が異なる場合、あるいは概念が更新された結果文字列が合致しない場合、nlm_yに対応するCUI・SUIを得ることはできない。そこでnlm_yに対応するCUI・SUIが得られないものについては医学中央雑誌で対応することとした。

日本語と英語との対応を確認する際に、ある用語・概念が所属するMeSH上の位置、すなわち概念間の階層関係を参照することができれば、誤りが少なくなり、作業効率も向上すると考えられる。また確認作業を分担して行うためには、各自の専門領域に沿った作業範囲を指定できる方が良い。そこで日本語医学用語を呈示する際にはMeSH Tree Number順に表示することができるようにした。そのためにMRSOのCODEからMeSH Tree Numberを取り出し、CUIとMeSH Tree Numberを対応付けるテーブルを作成し、利用した。

1.4. 作業用インターフェイス

日本語医学用語がUMLSに正しく対応していることを確認するためのインターフェイスをWeb上に作成した。作業班のホームページを図3に示す。作業を行う際にはユーザー認証を経て作業用のホームページに入る。作業方法のヘルプ画面はここで参照することができる。例を図4に示す。

確認作業は班員で分担して行うが、各自の専門分野に応じて好みの用語領域を選択できる方が良い。そこで各自の分担範囲を

MeSH Tree Number の最初の 1 桁で指定される作業エリアから幾つか選択することができるようにした。作業領域の選択画面を図 5 に示す。例えば B03 は「細菌、バクテリア」領域であり、この中に確認するべき概念が 442 個存在することがわかる。すでに作業者が割り当てられた領域はこのリストには表示されない。各班員は自分の割当分の作業が完了した後に、再度この画面に戻り、別の領域を自分で指定する。何も指定するものが無くなれば、全ての確認作業が完了したことになる。

確認作業を行うために 2 種類の画面を作成した。MeSH Tree Number 順に呈示するものと個々の概念ごとに呈示するものである。Tree Number 順に表示したものを図 6 に示す。画面は縦 3 つに分けられている。左欄に MeSH Tree Number と MeSH Heading、そしてそれに対応する日本語が表示されている。Tree Number をクリックするとこれが赤く表示され、対応する UMLS 情報が中央欄に、日本語情報が右欄に表示される。なお Heading が黒いものは未判定（日本語も黒い）、青く表示されているものは判定済みのものであり、判定が OK のものは日本語も青く、判定が NO のものは日本語が赤く表示される。図では MeSH Tree Number が A12 以下の部分が表示されているので、「液体・分泌物」に属する用語が Tree Number 順に呈示されていることがわかる。中央欄は UMLS 情報で、上から順に「CUI およびその典拠と定義」、「同じ CUI に属する英語表現一覧」、「同じ CUI の MeSH Tree Number とそのノード名称一覧」が呈示される。右欄は日本語情報である。まず判定用の OK および NO ボタンがあり、NO の場合にはその理由を記載するコメント欄がある。（OK の場合にもコメントを行うことはできる）。この下方に医学用語シソーラスの統制語が表示される。もし CUI に対応する ICD10 が存在し、さらにその ICD10 に対応する日本語病名が存在すれば、その病名が表示される。これは MEDIS-DC から配布されている「ICD10 対応電子カルテ用標準病名マスター」を利用している。この下に医学用語シソーラスの辞書と典拠が並ぶ。（dogi_no が 2 以上のものは参考情報として別に表示されている）。図では既に判定されたものであるため、判定結果とその理由も表示されている。

個々の概念ごとに判定作業を行う場合を図 7 に示す。これは図 6 から左欄を省いたものに近く、左欄が日本語情報で右欄が UMLS 情報となっている。内容は図 6 の時と同じである。この画面は割当領域の用語順に呈示されるものと、割当領域内の未判定の用語がランダムに呈示されるものとがある。

なお、これらのインターフェイスとコメントの付与方針などについてはワーキンググループで何度か検討を行い、詳細を詰めた。

C. 研究結果

1 結果

1.1 語彙数

UMLS 全体における CUI の数は 870,853 個である（英語のみに限っても同じ）。ちなみに登録されている用語数は 2,083,103 個であり、英語のみでは 1,753,789 個である。この中から MeSH に関連するものだけを抽出すると、CUI の数は 20,743 個、これに含まれる用語数は 179,856 個である。CUI

数で見るとUMLS全体の2%に過ぎないが、用語数で見ると約10%に相当する。

医学用語シソーラスに収載されている統制用語のうち、MeSHと対応しているものの数は19,180個であり、これに含まれる医学用語の数は同義語が26,598個、類義語が93,495個、関連語が32,260個であった。このうち、MeSH Headingを介してUMLSと対応を取ることのできたCUIの数は、19,133個であった。すなわちUMLS上でMeSHに属するCUI数である20,743から引いた残りの1,610個は日本語側に対応する概念が存在しないことになる。多くは2002年版との違い、および医学用語シソーラスに含まれていないZセグメントすなわち地理情報関連のものと思われ、対応を医中誌に依頼した。

1.2.確認作業結果

1月から3月までの2ヶ月で、対象CUI 207,43件のうち、7,794件で確認作業が完了、未判定のものは12,949件であった。作業に参加した人数は10名、1人当たりの判定数は最小800件、最大4,800件、平均1,800件であった。判定されたもののうち、OK判定数は7,215件、NO判定数は579件であり、何らかのコメントが付されたものは793件であった。これは確認数の約10%に相当する。コメント内容はオンラインでリアルタイムに参照することが可能であり、他の作業者がどのようなコメントを付しているのかを確認することができた。これを図8に示す。

コメントの内容で多数を示したものは以下の3件であった。

- 「日本語訳が存在しない」もしくは「英語表記は日本語として適当でない」とするもの：350件。
- MeSH Headingあるいは UMLS側で複数形表記となっているものに対して、日本語側に「～類とすべきでは」というもの：350件。
- 化学物質で「エステルも含むはず」というもの：53件。

これ以外のものが216件あり、その内訳は「日本語表記として不適なものがある」、「～という表記の方が適当」、「旧漢字での表記の方が良い」など日本語表記に関するもの、「上位あるいは下位概念が日本語に含まれている」や「～は異なる概念である」など概念として正確に対応しない用語が含まれている場合、の2種類に大別することができた。前者は日本語表記を修正する、もしくは表記を加える必要がある。後者はもとのMeSH（UMLS）の概念そのものに曖昧性が存在する場合が多く（例えば下位概念がMeSHに存在しないなど）、MeSHレベルの対応としては修正を要しないことになる。しかしMeSH以外の統制用語に対応するCUIが存在する可能性もあり、それをどのように検索するかについては今後の課題となった。

D. 考察

今回の作業が完了すると、意味上の曖昧さを排除した状態で英語と日本語との対応付けが実現する。これにより、この日本語文字列をバックボーンに据えることによって、日本医学会医学用語集とUMLSとの対応作業を有利に進めていく方策が考えられる。例えば医学用語集の日本語見出しと今回の作業で確定した医学用語を比較して一致したものを除く、あるいは一致したもの

で英語文字列を再検討することが考えられる。つまり従来は英語文字列だけを使って機械的一致を図ってきたものを、日本語文字列にも適用することができる。

日本語文字列に対しては、英語文字列のnormalize に相当する正規化手法が存在しない。しかし日本語医学用語の部分文字列から重要なものだけを取り出し、何らかの基準で不要なものを削除する、あるいは部分文字列を同義語で展開した組み合わせを作るなどの操作によって、英語のnormalize に相当する正規化を行うことができれば、微妙に異なるがために正しく対応付けられていなかった日本語文字列相互の比較が行えるようになる可能性はある。

コメント結果にもあるように、あるMeSH 概念に下位概念が存在するにもかかわらず、それがMeSH に定義されていない場合がある。その場合には対応する日本語も上位・下位概念が混在したものになり、曖昧性が発生する。しかし適当な下位概念はMeSH 以外の統制用語を探せば存在するかもしれない。つまりその曖昧なMeSH のCUI に対して child 関係または narrower relation を有するCUI をMRREL から探して来れば、その中に最適な候補が存在するはずである。このようにして、今回の作業で得られたコメント情報を利用することによって日英の対応をMeSH の外に拡張することができそうである。しかしその際にはMeSH とは異なる統制用語のCUI であることを知った上で応用する必要があることに注意しなくてはならない。

少なくとも、MeSH Heading に対応する信頼性の高い日本語医学用語のシソーラスが構築できたことは、ひとつの成果である。MeSH は文献検索に利用されている統制用語なので、その日本語側の語彙が対応することによって MEDLINE 検索に日本語キーワードを与えることが可能となる。さらに MRCOC を利用すればMeSH 概念相互の共起情報が得られるので、ある概念に関連する概念を検索することが可能となる。つまり通常の MEDLINE 検索で利用できるツール類をそのまま日本語でも利用できることになるので、医療従事者だけでなく一般の人々が容易に医学文献を検索できるようになる可能性がある。

同様に、今回の作業で得られた日本語文字列とCUI を介することによって、UMLS の様々なリソースを利用することができる。例えば複数の日本語医学用語間の意味的関係を調べるには、UMLS の semantic network を利用することができる。また今回の CUI はMeSH を典拠とするものに限られているが、MRREL を利用すれば他の統制用語集、例えばICPC、LOINC、Read Code、SNOMED International などとの対応を取ることが可能となり、応用範囲が広がる。またこれらの統制用語集を介することにより、UMLS には含まれていないSNOMED-CT などとの対応を取ることが可能となる。これを利用すれば、例えば「肝細胞癌」は「(英語で) 肝臓の悪性腫瘍である」という関係記述を得ることができる。もし個々の英語に対応する日本語が存在すれば、この関係記述を日本語で表記することができることになる。

Web インターフェイスに関して、本研究ではWeb ブラウザを利用して多人数が遠隔地から同時に確認作業を行うことのできる環境を提供し、その有用性を確認した。こ