紹介

統合医学用語システム (UMLS) 知識ソースの概要

川野惟二*

[著者抄録] 米国国立医学図書館 (NLM) が開発した統合医学用語システム (Unified Medical Language System : UMLS) 知識ソースについて、その概要を紹介した。UMLS 知識ソースの主要な構成部分であるメタシソーラス、語義ネットワーク、SPECIALIST 辞書、情報源マップについて、構成生物医学用語、その異形やそれらの用語間の関係に焦点を当て、また、それらデータの情報源などを含めて記述した。フォーマット、レコードなどソフトウェア関連事項についても簡単に触れた。

[著者付与キーワード] NLM, 統合医学用語システム, メタシソーラス, 語義ネットワーク, SPECIALIST 辞書, 情報源マップ, 生物医学用語, MeSH, データベース, 情報源, 知識ベース

Introductory article

The outline of Unified Medical Language System (UMLS) Knowledge Sources

KAWANO Koreni

[Author Abstract] The outline of Unified Medical Language System (UMLS) Knowledge Sources developed by the US National Library of Medicine (NLM) is introduced. Metathesaurus, Semantic Network, SPECIALIST Lexicon and Information Source Map which are main parts of UMLS Knowledge Sources are explained, focusing on biomedical terms, their variants and relationship between their terms, and including information sources of their vocabulary, databases, etc. Software-related matters such as formats, records, etc. are also briefly described.

[Keywords by Author] NLM, Unified Medical Language System, Metathesaurus, Semantic Network, SPECIALIST Lexicon, Information Source Map, biomedical terms, MeSH, databases, information sources, knowledge base

* メドラーズ・サーチ・アナリスト (元 JICST 特別情報員) (〒247-0053 鎌倉市今泉台6-4-1)
Medlars Search Analyst (4-1, Imaizumidai 6-chome, Kamakura-shi, 247-0053)

1. はじめに

本稿は米国国立医学図書館 (National Library of Medicine, 以下 NLM) が開発した統合医学用語システム (Unified Medical Language System, 以下 UMLS) 知識ソースについて、主としてその第 8 版ドキュメント (1997 年) により概要を紹介したものである。

UMLS 知識ソースの中核はメタシソーラス (Metathesaurus), 語義ネットワーク (Semantic Network), SPECIALIST 辞書 (SPECIALIST Lexicon), 情報源マップ (Information Source Map) であり, それらの中において, 構成する医学用語や用語間の関係などを中心に述べ, ソフトウェアなど関連する事柄についても簡単に触れた。

2. UMLS 知識ソースと UMLS プロジェクトの概要

UMLS 知識ソースは 1986 年より NLM が開発に着手し, その目的は保健関係従事者, 研究者の検索を助け, また, 各種情報源からの電子化した生物医学情報を統合化するシステムの開発を支援することにある。UMLS プロジェクトは用語の差異や多数のデータベースに分散する関連情報間に存在する検索問題を克服し, 広い範囲のアプリケーションプログラムにより利用可能な機械可読“知識ソース”の開発を行うことである。

その目標は計算機ベースの患者記録, 書誌データベース, ファクトデータベースやエキスパートシステムを含む, 本質的には異なる情報システムに利用者がリンクしやすくすることである。

それゆえ, UMLS へのアプローチは異なる利用者により表現される概念の差異を補い, 利用者の問い合わせに最も関連の深い情報源を識別し, そして, それらの情報源からの情報を検索するのに必要な通信および検索手順を通過するため, 多くのアプリケーションプログラムにより利用可能な機械可読“知識ソース”の開発を意味する。UMLS

知識ソースはまた, 保健データの作成および索引作成へのアプリケーションを支援することも可能である。

以下に述べる UMLS 知識ソースの内容は次の四つのものである。メタシソーラスは生物医学における概念, それらの各種名称とそれらの間の関係を含み, 語義ネットワークはメタシソーラス中において割り当てられた, すべての概念に対する一般的カテゴリーまたは語義タイプの一つのネットワークである。

SPECIALIST 辞書は生物医学用語についての統語的 (syntactic) 情報を含み, メタシソーラス中に現われる概念名称における構成用語を結果的にカバーするであろう。また, 情報源マップは機械可読生物医学情報の人間可読および機械処理可能な情報を含む。

UMLS プロジェクトは NLM 内部の研究開発チームが指揮し, そして契約と助成による外部研究グループを含んでいる。UMLS 関係の開発を行うため NLM により助成されたグループは Lexical Technology, Inc., Massachusetts General Hospital, Brigham & Women's Hospital, ピッツバーグ, コロンビア, スタンフォードの各大学, Mayo Clinic および Kaiser Permanente であるが, 世界で約 700 の機関や個人が各種のアプリケーションで UMLS 知識ソースを使って実験中である。

NLM 自体は MEDLINE 検索強化のため, Internet Grateful Med において UMLS メタシソーラスを使用している。

2.1 メタシソーラス

メタシソーラスは UMLS の中心的な語い (vocabulary) 構成物である。メタシソーラスの“meta”は“より包括的な, より優れる”の意に由来するもので, メタシソーラスは生物医学分野で使われる, 多くの異なる統制語い集や分類に出現する概念に関する情報のデータベースである。

メタシソーラスは意味, 階層関係, および, そのソース語い集の中に存在する用語間のそのほか

の関係を保持しているが、また一方では、その各概念についてのある基本的な情報を加え、そして、異なるソース語い集からの概念と用語の間の新しい関係を設定している。

2. 1. 1 メタシソーラスの語い

1997年メタシソーラスは30以上の語い集と分類からの33万以上の概念と739,439の異なる概念名称を含む。

メタシソーラスの語いは患者レコードシステムにおける使用のために設計した用語、すなわち、統計報告や請求書に対して使用する大きな疾病、処置の分類、精神医学、看護、医用装置、有害な薬物反応など、また、専門の診断システムからの疾病および所見用語や情報検索に使われる、いくつかのシソーラスの用語を含む。

メタシソーラスの構造は英語以外の言語へのソース語いの翻訳語を組み入れてあり、1997年メタシソーラスはNLMの医学件名リスト、Medical Subject Headings (以下MeSH)のフランス語、ドイツ語、ポルトガル語およびスペイン語の翻訳語を含む。

2. 1. 2 概念による組織化

メタシソーラスは概念または意味によって組織化され、本質的に、その目的は同じ概念の代替名称や見解をリンクすること、また、異なる概念の

間の有用な関係を識別することである。

メタシソーラスにおける各概念または意味はユニークな概念識別子(CUI)を持ち、また、メタシソーラスの各々のユニークな概念名称または文字列はユニークな文字列識別子(SUI)を持つ。大文字と小文字の違いも一つの別の文字列として扱われ、それぞれ別個のSUIを持つ。異なる言語(例えば、英語とスペイン語)における同じ文字列は各言語で異なるSUIを持つ。メタシソーラスでは英語のエントリーに対してのみ、各文字列はその辞書の異形または小さな変化のすべてに対して、共通の用語識別子(LUI)によってリンクされる。

それゆえ、メタシソーラスにおいては一つの英語用語は辞書の異形であるすべての文字列のグループであり、すべてのSUIおよびLUIは少なくとも一つのCUIにリンクされ、また、同じ意味を持つ異なる用語は同じCUIにリンクされる。

これを図1で説明すると、文字列“Atrial Fibrillation”(心房細動)とその複数形“Atrial Fibrillations”は異なるSUI(S0000042とS0000043)を持つが、同じLUI(L0000032)にリンクされる。同様に、“Auricular Fibrillation”(心房細動)とその複数形“Auricular Fibrillations”はまた異なるLUI(L0000150)にリンクされる。Atrial FibrillationとAuricular Fibrillationsは

図1 概念、用語、文字列の関連

同じ意味を持つと判断されるので、それらの異なる LUI (L0000032と L0000150) は同じ CUI (C0000015) にリンクされる。

さらに、概念名称間に優先順序がある。それぞれの語い集に preferred terms (優先語) があってよく、例えば、図 1 において、Atrial Fibrillation は優先名称としてラベルされている。

また、代替用語との間の困難な選択を避けるために、同じ語い集からの文字列の異なる同義語、例えば、優先語、相互参照語、省略形はこの順序で異なる地位を持つであろう。順序の設定で考慮される要素としては主題カバレッジの広さ、更新頻度、その概念名称が通常の臨床または生物医学論文で使われる程度による。

2. 1. 3 異なる語い集における同じ文字列で異なる概念の扱い

相対的に少ないが、通常、メタシソーラスの中の異なる語い集の間において、同じ文字列が異なる概念である場合がある。図 2 でこれを説明する。

文字列“Cold”は MeSH では温度の度合いに対する名称であるが、SNOMED II では Cold は“Common Cold” (風邪) に対する代替名称である。結果として、文字列 Cold は異なった SUI を持ち、それぞれ異なった概念にリンクしている。そして、メタシソーラスにおいて、Cold に対する複

数の意味を機械可読の形で表現するため、Cold <1> と Cold <2> が創出される。

2. 1. 4 異なる概念間の関係

メタシソーラスでは異なる概念間の関係をいくつかのタイプで表現する。多くの関係がソース語い集から直接導出される。例えば、“Atrial Fibrillation”と“Arrhythmia” (不整脈) との間に一つの関係があるという事実は MeSH 階層ツリー構造から導出される。Atrial Fibrillation が Arrhythmia の一つであるという関係の正確な性質は MeSH には表現されないが、その関係はメタシソーラス構築中に加えられた。

異なるソース語い集からの概念間の関係もメタシソーラス構築中に作成されている。例えば、CO-STAR の中の概念“Proxysmal Atrial Fibrillation” (発作性心房細動) はメタシソーラスにおいて、Atrial Fibrillation へのより狭義の関係を持つとして識別される。

2. 1. 5 概念、用語および文字列の属性

メタシソーラスにおける情報のそのほかの部分は概念の属性、用語の属性または文字列の属性のいずれかである。例えば、語義タイプ“Disease or Syndrome” (疾病または症候群) とその定義“Disorder of Cardiac rhythm characterized by ……” (……によって特性化した心臓性律動不全)

図 2 語い集が異なるため同じ文字列が異なる概念を持つ例

は優先名称 Atrial Fibrillation の持つ概念の属性である。

以上、メタシソーラスとその構成用語、および、それらの用語間の関係の主なものを述べた。メタシソーラスは ASCII リレーショナルフォーマットと ASN.1 (Abstract Syntax Notation.1) フォーマットにより配布される。

2. 2 語義ネットワーク

語義ネットワークは UMLS メタシソーラスの中のすべての概念に対して一貫したカテゴリとそれらの概念間の重要な内部関係を提供する。1997年版の語義ネットワークは135の語義タイプと51の関係を含む。語義ネットワークはメタシソーラスの概念に対して割り当てられるが、それはまた情報源マップの中のデータベースに割り当てられる語義タイプに対するオーソリティとしても役立つ。

2. 2. 1 語義タイプとそれらの間の関係

語義タイプはネットワークにおけるノードであり、また、それらの間の関係はリンクとなっている。生物、解剖学的構造、生物的功能、薬物などに対する語義タイプの大きなグループがある。語義タイプの現在の適用範囲は広く、それは複数の領域における用語の広い範囲の語義カテゴリー化であり、また、主要なリンクは“isa” (…は…の一つ：筆者注) リンクである。これはネットワーク内のタイプの階層を設定し、そして、一つのメタシソーラス概念の割り当てに対して利用可能な最も特定の語義タイプの決定に使用される。

さらに、タイプ間の潜在的に有用な非階層関係のセットが確認されている。それらは “spatially related to” (空間的に関係する), “temporally related to” (時間的に関係する), “functionally related to” (機能的に関係する), “conceptually related to” (概念的に関係する) などの五つである。

2. 2. 2 関係の継承

可能な場合にはいつでもネットワークにおける高いレベルのノード間の関係が指定され、そして、

1. Organism
 1. Plant
 1. Alga
 2. Fungus
 3. Virus
 4. Rickettsia or Chlamydia
 5. Bacterium
 6. Animal
 1. Invertebrate
 2. Vertebrate
 1. Amphibian
 2. Bird
 3. Fish
 4. Reptile
 5. Mammal
 1. Human
2. Anatomical Structure
 1. Embryonic Structure
 2. Anatomical Abnormality

図3 語義タイプの階層の一部

その関係は通常それらのノードのすべての直下位ノードによって “isa” リンクを通じて引き継がれる。例えば、関係 “Process of” (…の過程) は語義タイプ “Biologic Function” (生物的功能) と “Organisms” (生物) との間の関係の保持が決まると、それはまた生理学的機能や生物的功能である “Organ or Tissue Function” (器官または組織の機能) と生物である “Animal” (動物) との間の関係を保持する (図3 参照)。

このように、関係は語義タイプの間で決められるが、しかし、それは必ずしも語義タイプに指定されている概念のすべての場合に適用されるものではない。関係は概念の特定のペアの間で保持されるかも知れないし、保持されないかも知れない。ある場合にはネットワークにおける語義タイプと引き継がれるべきリンクとの間に矛盾があるであろう。もしそうであればリンクの継承はブロックされる。

例えば、引き継ぎによって語義タイプ “Mental Process” (精神的過程) が “Plants” (植物) の “Process of” であるとする。植物は意識のある生物ではないので、このリンクは明らかにブロッ

クされる。

MeSH 階層の部分においても、階層的に直上位の語義関係に対して直下位をラベルしており、これは生物科学部分と同様、解剖学、疾病、精神医学および心理学の各セクションもラベルされている。語義ネットワークにおける“isa”関係とすべての非階層関係はこのラベル付けに対してバリッドであり、少数の例外を除いて MeSH 用語間のリンク関係は語義ネットワークにおいて反映される。すなわち、もし二つの MeSH 用語がある関係によってリンクされるならば、そのリンクはそれらの MeSH 用語に割り当てた語義タイプ間のリンクとしてネットワークにおいて表現される。

2. 3 SPECIALIST 辞書

この辞書は SPECIALIST 自然言語処理システム (Natural Language Processing System, 以下 NLP) に対して必要な辞書情報を提供するために開発され、それは多数の生物医学用語を含む一般の英語辞書であることを意図している。その範囲は一般に発生する英語の語および生物医学語を含み、各語または各用語に対する辞書エントリは SPECIALIST NLP システムに必要な統語上、形態上などの情報を含む。

取り扱い上、辞書異形生成プログラムの課題は自然語および用語における高度な変化の程度にあり、語は同じ語であっても適切に考慮されるべき語尾変化形がある。例えば、動詞“treat” (加療する、など) は三つの追加語尾変化形を持つ。すなわち、第三人称単数現在形の“treats”、過去および過去分詞形の“treated”と現在分詞形の“treating”である。

2. 3. 1 辞書の構成

辞書は各スペルまたは言葉の特別な部分におけるスペル変化に対するエントリを持つ一組の辞書エントリから構成される。辞書項目は、複合語が一般英語または医学辞書において、または MeSH のような医学シソーラスにおいて一つの用語として存在し、そして辞書項目であると決められるならば、その他の語からなる複合語として

扱われる。一般に使われる頭字語 (acronym) または省略形も複合語として許される。MeSH における LMWH は Low-Molecular-Weight Heparin (低分子重ヘパリン) の頭字語の例であろう。

基本形とスペル変化を与えるエントリはユニットフォーマットにおいて一つの辞書レコードに収容される。基本形はスペル変化のうちの一つの引用 (citation) 形であり、その引用形は動詞の場合は不定詞、名詞の場合は単数形、語尾変化する形容詞の場合は原級である。

ユニットレコードはスロットとフィラーで構成する一つの枠構造となっている。図 4 のように、一つの辞書レコードはカッコ {…} で区切られている。図 4 の場合、基本形 base=anaesthetic (麻酔薬) で、任意にスペル異形 spelling variants=anesthetic を持っており、さらに、entry=E0008769 としてエントリ番号 (EUI) を示し、“E”を先頭に持った 7 桁の数字が続く。各エントリは語のカテゴリーを示してあり、cat=noun (名詞) はこの用語が名詞であることを、また、異形態を指示する異形 variants=reg はこの名詞がカウントできる普通の (regular) 名詞で英語複数形 (この場合、anaesthetics) を持つことを示している。

さらに、このレコードは形容詞エントリ cat=adj として形容形 (adjective) があることを、また、variants=inv はこの形容詞は比較級、最上級を作らないことを指示している。最後の position=attrib はこの形容詞が名詞修飾語 (attribute) であることを指示している。

```
{base=anaesthetic
 spelling variant=anesthetic
 entry=E0008769
   cat=noun
   variants=reg
 entry=E0008770
   cat=adj
   variants=inv
   position=attrib(3)
}
```

図 4 anaesthetic の辞書レコードの例

2. 3. 2 意味で分割できない辞書エントリー

辞書エントリーは意味によって分割できない。名詞“act”は頭文字を大文字化したスペルと小文字スペルとがあり、それぞれの意味を持っており、また、an act は一つの“行為”と“法令”の意味を持っている。しかし、それらの意味は同じスペルと統語上のカテゴリーを与えられ、現在の辞書では単一の辞書エントリーによって表わされている。

```
{base=Act
  spelling variant=act
  entry=E0000154
    cat=noun
    variants=reg
}
```

図5 Act の辞書レコードの例

“Act”に対するユニットレコードを図5に示した。この場合、動詞の act は頭文字を大文字化したスペルを持っていないので、この辞書レコードには含まれない。

また、異なった意味が異なった統語上の挙動を持つときも、それぞれに対するコードが単一のエントリーにおいてレコードされる。例えば、“beer”は二つの意味を持っている。すなわち、アルコール飲料と飲料の標準容器の量（ビールの本数）である。次の文のように

- 1) ビール (beer) を飲んだ患者はワインを飲んだ患者より早く回復した。
- 2) 36人の患者は1日5本以上のビール (beers) を飲むと報告した。

```
{base=beer
  entry=E0012226
    cat=noun
    variants=uncount
    variants=reg
}
```

図6 beer の辞書レコードの例

1の意味はカウントできない物質名詞、2の意味は普通名詞である。図6のように、このような両方の意味に対する適切なコードがエントリー中に含まれる。

2. 3. 3 辞書の情報源の範囲

辞書のコード化に対して、多数の情報源からの語が選択されている。メタシソーラスと Dorland's Illustrated Medical Dictionary の双方に出てくる語とともに、MEDLINE の抄録の UMLS Test Collection からの約2万語が SPECIALIST 辞書の核心を形成する。

さらに、一般の英語語彙集からの語を含むよう努力がなされている。American Heritage Word Frequency Book からの最も頻度の高い1万語と Longman's Dictionary of Contemporary Dictionary において定義に使われている約2万語も入力されている。

コード化される語の大多数が名詞であるので、現在の MEDLINE の文献 (citation) レコードの中の動詞を識別することと Computer Usable Oxford Advanced Learner's Dictionary の使用により、また、McCray & Srinivasan が開発した自発研究法で前記 Dorland 医学辞書からの可能性ある形容詞を識別して、動詞と形容詞の追加への努力を行っている。これらのほかにも多くの辞書などがコード化に使用されている。また、各データエレメントは ASCII リレーショナルフォーマットにおいて、フィールドとして表現されている。

2. 4 情報源マップ

情報源マップの長期的目標は利用者が生物医学情報に対する質問を課した場合、その質問に最も関係ありそうな情報源を決め、識別した情報源の範囲、アクセス条件などについて人間可読情報を提供し、識別した情報源に自動的に接続して適切な情報を検索することである。

情報源マッププロジェクトは主要な二つの部分からなり、一つは生物医学情報源を記述したレコードを含むデータベースであり、もう一つはそ

のデータベースを管理および使用する WWW ベースの一組のアプリケーションである。

2. 4. 1 情報源マップデータベースの情報源

情報源マップは現在、15の情報プロバイダと70の情報源を含むが、そのうち53は NLM から、17は非 NLM ソースにより提供される。また、その情報源は生物医学研究、臨床処置、生物倫理をカバーする書誌データベースだけでなく、AI/RHE-UM (NLM)、DxPlain (Massachusetts General Hospital)、Iliad (ユタ大学) および Quick Medical Reference (QMR, CAMDAT Corp.) のような診断エキスパートシステム、History of Medicine (NLM) からのオンライン画像などを含む画像データベース、遺伝、蛋白質および核酸配列、薬物、毒物、環境保健関係のファクトデータベースを含む幅広いものである。

2. 4. 2 情報源マップデータベースレコードのエレメント

情報源マップデータベースのデータベースレコードは単一の情報源を記述し、レコードは各種のフリーテキスト記述、関連する資源の内容の性質、情報源の起源や規定などについての情報を記述した、予め定義したバリューを持つ明確なデータを含むものである。

各情報源の概念的な範囲は MeSH 用語と MeSH の副標目 (subheadings) を使って索引される。MeSH の用語を使って生物医学文献を索引するときは可能な限りスペシフィックな仕方で行うが、それとは対照的に情報源マップレコードの情報資源全体に適用する索引語はその情報資源の性質を一般性の最も高いレベルで特定化する。

情報源マップレコードは語義ネットワークの語義タイプと語義タイプとの間の関係を使っても索引される。語義タイプの関係は一つの動詞 (関係付けるもの) によって接続された二つの名詞 (語義タイプ) として考えることができる。例えば Hazardous or Poisonous Substance causes Pathogenic Function (有害または毒性物質は病原性機能を引き起す) というように、毒性データベースの関係の記述に使われる。このアプローチ

は自然語の能力を持ったソフトウェアツールの開発を促進するかも知れない。

2. 4. 3 情報源マップアプリケーション:

Sourcerer と Apprentice

適切な情報源から自動的に情報を検索するために情報源マップデータベースを使うプロトタイプツールである Sourcerer はいくつかの国際会議でもデモンストレーションを行っているが、Sourcerer はさらに発展しつつある。WWW を利用したネットワークベースの複数の操作台、マルチプロトコル、マルチメディアに基づいて、Sourcerer は技術的に健全でオープンに利用可能な、注意深く選定した通信およびソフトウェア標準に基づいており、それは NLM が開発中の情報源マップツールの目標を反映している。

一方、WWW ベースの Apprentice も開発中である。Apprentice は情報源マップデータベースに包含される可能性のある生物医学情報資源を情報プロバイダに記述させるのを可能ならしめるであろう電子的様式を作り出す。

情報源マッププロジェクトは NLM が研修した索引者の助力により、要求されるように各プロバイダが増大する情報源マップレコードを適切に作成し維持する、分散したメンテナンス環境を創造することを意図している。さらに開発とテストが完全になったのちに、NLM は共有の生物医学情報源マップの構築における広範な参加を探索するであろう。

5. お わ り に

以上、UMLS 知識ソースの主要な構成部分について、主として、第8版ドキュメント (1997) により、構成用語や概念、それらの間の関係を中心に紹介し、関連するソフトウェアなどについても簡単に触れた。UMLS 知識ソースは先端的な知識ベースの内容を含むと思われる、一部に十分に理解し難い点もあったが、本稿全体として UMLS 知識ソースの概要を理解いただければ幸いである。

本稿作成に当たって、科学技術振興事業団情報

提供部細江孝雄オンライン課長より、親切な執筆のお勧めをいただき、また、同部諏訪秀策特別情報員により、インターネットを通じて膨大なド

キュメントを出力していただいた。両氏に対し、深甚な謝意を表する。

~~~~~ 参 考 文 献 ~~~~~

- 1) UMLS Knowledge Sources : 8th Edition, January 1997 Documentation
- 2) UMLS Knowledge Sources : 4th Experimental Edition, April 1993 Documentation
- 3) NLM Fact Sheet : Unified Medical Language System (<http://www.nlm.nih.gov/pubs/factsheets/nmls.html>)
- 4) 渡辺俊彦, 福島勲, Internet Grateful Med と統合医学用語システムの概要. 情報管理. 39(4) 267-283 (1996)