ライフサイエンス辞書からクリニカル インフォマティクスへ 臨床テキストからの知識発見に向けて

From a life science dictionary to clinical informatics
Toward the knowledge finding from medical texts

金子 周司¹ | 大武 博²

KANEKO Shuji¹; OHTAKE Hiroshi²

.....

1 京都大学大学院薬学研究科 (〒606-8501 京都府京都市左京区吉田下阿達町46-29) Tel : 075-753-4541

E-mail : skaneko@pharm.kyoto-u.ac.jp

2 福井県立大学学術教養センター (〒910-1195 福井県吉田郡永平寺町松岡兼定島4-1-1)

1 Graduate School of Pharmaceutical Sciences, Kyoto University (46-29 Yoshida-Shimoadachi-cho Sakyo-ku Kyoto-shi, Kyoto 606-8501)

2 Center for Arts and Sciences, Fukui Prefectural University (4-1-1 Matsuokakenjojima Eiheiji-cho Yoshida-gun, Fukui 910-1195)

原稿受理 (2010-10-05)

情報管理 53(9), 473-479, doi: 10.1241/johokanri.53.473 (<http://dx.doi.org/10.1241/johokanri.53.473>)

著者抄録

生命科学領域の研究報告で用いられる英語, 日本語の専門用語を各々 10 万語規模で収集して対訳づけを行い, 類義語と階層構造の整理を行った対訳シソーラスを完成させた。続いてこれを利用して, テキスト中での特徴的な用語の共起頻度に基づく関連性を求めることで専門知識の検索に有用な連想検索を試作し, これらを Web 辞書として公開した。また, 新たに読解支援ツールとして複合語に対応したマウスオーバー辞書を開発, 公開した。さらに, 医療文書からの知識発見に資することを目的として, 米国 FDA が公開する有害事象データベースから医薬品名の名前解決に取り組んだ。これらの試みについて紹介する。

キーワード

医学シソーラス, MeSH, 連想検索, 有害事象報告システム, テキストマイニング, マウスオーバー辞書

1. はじめに

ライフサイエンス辞書 (Life Science Dictionary: LSD) は生命科学・医学関連領域で用いられる英語および日本語の専門用語を, 国内外の文献抄録, 教科書, 総説等から集めたテキストコーパスでの頻度解析に

基づいて収録し, 対訳関係を定義した英日双方向の電子辞書である¹⁾。われわれは用語の補充を続けるとともに, LSDに収録された概念の整理 (シソーラス化) を始めている。このシソーラスの目的は大きく2つあり, その1つは日本人による情報検索を手助けするような使い道²⁾, もう1つは医療情報のデータマイニン

グ等による知識抽出のための参照辞書としての応用³⁾である。本稿では、現在までに公開あるいは発表した試みを紹介したい。

2. Web辞書のシソーラス拡張と連想検索

インターネットを介して得られる医療健康情報は、一般市民だけでなく医療従事者および医療系学生にも影響を及ぼしつつある。特に、日本語で記述された解説記事などの増加は、相対的に英語で記述された原著論文、総説、診療ガイドライン、規制情報など、科学的に評価が高くグローバルに流通している科学情報が利用されなくなる状況を招いている。しかし、もし英語のリソースを日本語で検索できるポータルや、Webで公開される英文を容易に読解するための支援ツールが提供されれば、利便性向上と有益な波及効果が期待できる。そこでわれわれは、インターネットで公表される広範な医学関連研究の成果を、日本人に検索あるいは理解しやすくすることを目的として、過去10年以上にわたって構築してきたLSD資源を利用するシソーラスの制作と日本語ポータルを開発し、さらにマウスオーバー辞書を制作した。

2.1 シソーラス制作

医療情報の理解に必要と考えられる専門用語の異表記を統一するための統制語としては、後で情報検索に利用することを考えた結果、米国国立医学図書館 (National Library of Medicine) が構築しているMeSH (Medical Subject Headings) に準拠することにした⁴⁾。そこでMeSH 2010 (2009年11月版) より解剖部位 (A01-A17)、生物名 (B01-B08)、病名および症候名 (C01-C23およびF03)、生体分子および医薬品名 (D01-D27およびSupplemental Concepts)、方法および尺度 (E01-E07)、学問領域や現象 (G01-G14)、専門分野 (H01-H02) に帰属する専門用語から、上記のカテゴリに帰属するDescriptor 23,047語とLSDに同一の見出し語が収録されていたSupplemental

Concepts 4,515語を合わせた計27,562語を統制語として採用した。これを元に、LSDとMeSHのすり合わせ作業を行い、19万7,000語の英日シノニム (同義語) 辞書を制作した。なお一部の統制語においてはMeSHに準拠せず、出現頻度の高い語句を採用した。

2.2 連想検索へのシソーラス応用

連想検索とは、検索語を入力する際に、その検索語に関連する別のキーワードを自動的に表示して、関連知識との掛け合わせによる情報の絞り込みを簡便に行えるシステムである。連想検索で提示する用語はコーパスの共起解析によって得られる⁵⁾ が、現在までに公開されている連想検索⁶⁾⁻⁸⁾ での出力を調査した結果、同義語の重複や過分割による不要語の混入が指摘できた。そこで、国内外の研究情報を参照するような専門家の利用を想定したLSDにとっては、既存の形態素解析辞書に依らず、自作のシソーラス辞書を用いてPubMed英文抄録における共起解析を行うことで最適化された関連ワードのセットが得られると考えた。

2.3 用語の共起解析

解析対象のコーパスとして1995年以降、インパクトファクターの高い100誌に掲載された欧米の研究機関からの抄録43万論文 (8,000万ワード、550MB) を用いた。また、解析辞書としてはLSDシソーラスを用い、自作Perlスクリプトによってコーパス中で最長一致する専門用語に統制語のXMLタグを日本語表記で施した。

続いてXMLタグの出現頻度と1抄録単位での共起頻度を計数し、tf-idf⁹⁾ によって関連性を算出した。この関連語の内容を目視によって点検し、周知と思われる知識、あるいは不適切にタグづけられた用語をマークし、解析辞書から除去した。削除された用語としては、「酸」「手」など1文字で表される用語、および「細胞」「ラット」「タンパク質」等、高頻度で全論文の数%以上に出現し、tf-idf補正をもってして

もランクインするような基本語（計約120語）であった。

このようにして辞書を改訂した後、再解析を行った。結果としてコーパス中に1回以上出現した統制語は22,931語（83%）、検出された共起関係は336万組であった。得られた共起関係が多数であったため、tf-idf値によってソートした1語あたり最大30語を各統制語について付与する関連ワードとした。

2.4 公開Web辞書への実装

以上のようにして制作したシソーラスおよび連想

検索は、従来から無償提供しているインターネット電子辞書WebLSD¹⁰⁾に付随する形で公開した。収録語数は2010年1月版において、英和10万3,143語、和英11万6,740語、用例22,261文、音声16,146語であり、シソーラスツリーは同じ画面でシノニムや連想検索語とともに表示できるようにした（図1）。また、標準病名と一致する表示についてはMEDIS v2.82に準拠したICD-10コードを付与した。

連想検索として提示した関連語の内容については図1の例を用いて考察する。この例では「睡眠時無呼吸症候群」という疾患の原因となる「肥満」や併発する「傾眠症」、「いびき」といった症状の他、診断法である「睡眠ポリグラフ」や治療法である「持続的気道陽圧法」が上位にランクされ、適切な専門知識が抽出されているようである。他の見出し語についても辞典における見出し語の解説で用いられるような重要な専門用語が上位にランクされ、直接の関連性を示しているシソーラスでは得られない有意な関係が数多く提示できた。

また、この例においては中位以下に「エストロゲン補充療法」や「プロゲステロン受容体」など、一見、睡眠とは無関係と思われる専門用語が提示されている。しかし、これらはリンク先の専門文書を読むことによって、閉経期にある女性に特有な女性ホルモン低下と睡眠時無呼吸の関連について最新の研究が進んでいることが読み取れる。同様に、他の見出し語についても疾患と治療薬の関係や遺伝子と疾患の関連など、新しい知識へのポータルとして専門家に対しても有用な関連語が選択できたと思われる。

しかしながら、概念ツリーで近接する不要な用語あるいは検索がヒットしない用語の混入など、いくつかの改善を要する点も見出されており、今後、使いやすいインターフェースへの改良も含めて改善を重ねていく予定である。

英和・和英辞書からのリンクをたどることによって、同義語、概念ツリー、連想検索語が表示される。連想検索語からは、Google（日本語の場合）やEntrez（英語の場合）に外部リンクしている。

図1 WebLSDに実装したシソーラスと連想検索

2.5 マウスオーバー辞書の制作

以上の連想検索によって、日本語による英文情報

の検索を支援する方策が提案できた。一方、Webで提供される英語情報の読解についても効率を高めるツールの必要性があった。そこで、われわれはパソコンのプラットフォームに依存しない無料WebブラウザであるFirefox3のアドオンとして利用できるマウスオーバー辞書を開発した。

本ツールには、複数の単語で構成される専門用語も最長一致により最適な訳語を表示する機能を、わが国で初めて実装した。また、動詞や形容詞など用語の語尾変化も自動で解釈されるようにした(図2)。

本辞書ツールにより、ブラウザで表示される英文で意味を調べたい語句上にマウスカーソルをかざすだけで、10万語を収録したライフサイエンス辞書の英和参照が可能になった。本ツールは無償で公開している¹¹⁾。このマウスオーバー辞書から連想検索への発展検索については、現在、実装方法を検討中である。

3. 医薬品名の名前解決

米国FDAのAdverse Event Reporting System (AERS)¹²⁾は、2004年初頭からの数百万件の薬物有害事象に関わるレポートを公開している。各レポートの適応症と有害事象(すなわち病名や症状)の記述にはICH国際医薬用語集MedDRA¹³⁾のPreferred Term (PT)が統一的に用いられているため、適応症や有害事象ごとの集計が可能でありデータマイニングの格好の題材となりうる。しかし、最大の問題点はレポート中に医薬品名として一般名だけでなく、世界中で使用されている医薬品の商品名に加えて、注釈や用量など雑多な表記が含まれ、まったく正規化されていないことである。そこで本研究では制作したシソーラスの応用性を確かめ、データマイニングへの応用を探る目的で、AERSに記載された膨大な医薬品名の解決を試みた。

Webページ中の生命科学用語の上にマウスカーソルをかざした際に英和訳語が表示される。カーソル前後で最長一致する専門用語を表示する点に特徴がある。ショートカットキーとの組み合わせでオンデマンドで呼び起こすこともできる。

図2 ライフサイエンス辞書ツール(マウスオーバー辞書)

3.1 医薬品名の整理

FDAサイトで公開されている2004年から2008年第2四半期まで4年半分のAERSレポートをFileMaker Proを用いてリレーショナルデータベースに再構築した。適応症や反応事象の解決にはMedDRA/J ver.11.0のPT(あるいはLLT)とのリンクを設けた。LSDおよび収集した各種の医薬品名辞書についても、同様にリレーショナルデータベースに再構築して互いを名称でリンクさせた。

重複を考慮しない150万件のAERSレコードに記載された医薬品レコードは555万件であり、同一文字列ごとに整理すると20万867種類の異なる表記が含まれていた。これらの出現頻度を計数すると、一般名や商品名が正しく表記されている高頻度のレコードに混じって、医薬品名に用量、剤形、注記や意味不明のコード番号などを付記した低頻度のレコードが極めて多種多様に混入していることが明らかとなった。

3.2 医薬品名解決辞書 の制作

LSDシソーラスを基本として、MeSH Termに収録された欧米での異表記を統合することによって、同一の薬理活性成分を含む場合を同義語辞書に統合する作業に着手した。その際、塩および水和物は活性

本体と同等であると見なして統合した（例、塩酸モルヒネと硫酸モルヒネはモルヒネに統合）。しかし、体内動態が大きく異なることが予想されるエステル化ステロイド、プロドラッグ、包接化合物などは別化合物と見なして独自の統制語を立てた。また、増加しつつある合剤に関しても別と見なして、含有成分を列挙した統制語に統一した（例、アバカビル・ラミブジン・ジドブジン合剤）。さらに医薬品名を MeSH Pharmacological Actionsに関連づけ、各薬物に薬効分類および化学構造タグに関連テーブルの形で付与した。この作業の結果、日本語および英語の医薬品名45,000種類を収集し、約3,800種類の統制語に定義づけた。

ここでAERSに記載されていないながら統制語に収録されていない医薬品が800種類ほど残された。その多くは漢方薬および日本国内でのみ承認されている薬理活性成分であり、これらを補充するためシソーラスを拡張し、最終的に20万種類の医薬品名のうち、13万種類を特定できるAERS医薬品名解決辞書を制作した（図3）。

3.3 名前解決

この辞書を用いることによって、AERSに出現する医薬品レコードの97%（主たる原因と考えられる Primary Substanceに限定すると98%）が同定できた。また、すべての医薬品について、先に述べたように分子作用点、薬効分類、構造分類の属性が付与されているため、例えば、「ベンゾジアゼピン系催眠薬」、「フェノチアジン化合物」、といった、ある一群の医薬品を指す用語を用いてもAERS症例の検索が日本語で可能になった（図4）。

4. おわりに

大量のデータは、新たな学問領域を創出するポテンシャルを秘めている。事実、過去に遺伝子情報の蓄積はDNA塩基配列から生物学的に意味のある情報

AERS中には20万種類の医薬品名が存在し、その頻度分布は両対数逆比例となった。LSDシソーラスに収録された同義語のみで解決できる語の頻度分布（Assigned）を○で、修飾語を無視した医薬品名解決辞書を用いて解決できる割合（Resolved）を▲で示した。

図3 LSDシソーラスによるAERS医薬品名解決

を発見するbioinformaticsを生み出した。同様に今後、各種の医療文書がデータとして電子化されることは、その中から意味のある知識を抽出するclinical informaticsとでも呼ぶべき学問を生み出すだろう。しかし実際の医療記録に用いられる用語はさまざまであり、必ずしも統一されていない。したがって、clinical informaticsを興すために最も重要なことは、用語法（terminology）の解決である。

clinical informaticsは医療の安全に寄与できる。医療情報化社会において医療等の安全を達成するためには、市販後の医療情報や調査データを解析することによる有害作用の知識発見を早期に、確実かつ網羅的に行う必要がある。それに加えて、大量に蓄積された有害作用情報から、さまざまな統計学的な解析を可能にする薬剤疫学を発達させることは、基礎研究へのフィードバックを含めて学術的に興味ある課題である。すなわち、抽出される薬物（A）と有害事象（B）との関係抽出にとどまらず、DNAアレイを用いた他の研究における薬物（A）と遺伝子（C）との関係と論理的に組み合わせることによって、有害事象（B）と遺伝子（C）との関係が示唆されることになり、副作用メカニズムを実験科学で立証するための着眼点が提案されうるだろう。これらに通じ

(上)「MAO阻害薬」と「ベンゾジアゼピン系薬物」の併用例を検索した場合。(下)「ビスホスホネート剤」と「NSAIDs」の併用例を検索した結果。いずれの検索語も、シソーラスによって統制語に置き換えられ、該当する薬物がすべて検索対象となる。なお、薬理情報は一般的に1薬物に複数個定義されており、画面には1つしか表示されていない。

図4 AERS医薬品名の解決例

る第一歩として、AERSを題材に用いる場合に最大のネックとなる医薬品名をシソーラスによって解決できたのは意義深い。

AERSデータは適応症と有害事象の記述は正規化されている一方、医薬品名は報告者の入力のままに公開されており、膨大な種類の表記が存在していた。この解決に既存の医薬品集を統合したシソーラスを元にして拡張した医薬品名辞書が97%の適合率で有効だけでなく、薬理作用や構造情報とのリンクを可能にしたことは、自由度の高い検索を可能にするとともに、将来的に他の情報資源との直接リンクを可能にするだろう。今後、既存の医薬品に対する網

羅性を高めるとともに、新しく市場に登場してくる医薬品について随時フォローアップの努力を続けていきたい。

謝辞

本研究は、日本学術振興会科学研究費補助金（研究成果公開促進費 2006-2009）、同特定研究「応用ゲノム」（2008-2009）、厚生労働省科学研究費（地域医療基盤開発推進研究事業2006-2008）、JSTシーズ発掘試験研究（2009）の助成を得て行った。LSDプロジェクトの他のメンバー（藤田信之、河本健、鵜川義弘、竹腰正隆、竹内浩昭）の多大なる協力に感謝する。

参考文献

- 1) ライフサイエンス辞書プロジェクト. <http://lsd.pharm.kyoto-u.ac.jp/>, (accessed 2010-09-29).
- 2) 金子周司, 鵜川義弘, 大武博, 河本健, 竹内浩昭, 竹腰正隆, 天野博夫, 藤田信之. “医学用語シソーラスに基づく

く効率的医療情報検索システムの開発”. 医療情報学. vol. 28 (Suppl.), 2008, p. 639-642.

- 3) 金子周司. 薬物有害事象症例AERSの医薬品名解決と薬物分類および化合物構造からの検索システム. 医薬ジャーナル. 2010, vol. 46, no. 1, p. 125-132.
- 4) 金子周司. ライフサイエンス辞書とは. 情報管理. 2006, vol. 49, no. 1, p. 24-35.
- 5) 長尾真, 佐藤理史編. 自然言語処理. 岩波書店, 1996, p. 387-391, (岩波講座ソフトウェア科学15).
- 6) 国立情報学研究所. Webcat Plus. <http://webcatplus.nii.ac.jp>, (accessed 2010-09-29).
- 7) Preferred Infrastructure. “連想検索エンジン reflexa”. Labs.preferred.jp. <http://labs.preferred.jp/reflexa>, (accessed 2010-09-29).
- 8) National Center for Biotechnology Information. <http://www.ncbi.nlm.nih.gov>, (accessed 2010-09-29).
- 9) 長尾真, 佐藤理史編. 自然言語処理. 岩波書店, 1996, p. 419-421, (岩波講座ソフトウェア科学15).
- 10) ライフサイエンス辞書プロジェクト. “WebLSD”. ライフサイエンス辞書プロジェクト. <http://lsd.pharm.kyoto-u.ac.jp/ja/service/weblsd/index.html>, (accessed 2010-09-29).
- 11) ライフサイエンス辞書プロジェクト. ライフサイエンス辞書ツール. <http://www.lsdtools.org/>, (accessed 2010-09-29).
- 12) “Adverse Event Reporting System (AERS)”. U.S. Food and Drug Administration. <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm>, (accessed 2010-09-29).
- 13) 日本公定書協会. “MedDRA/J”. 日本公定書協会. http://www.sjp.jp/~jmo_new2006/php/indexj.php, (accessed 2010-09-29).

Author Abstract

We collected more than one hundred thousand English and Japanese terms respectively from medical research reports and compiled a bilingual life science thesaurus with synonym table and tree structure. We have published a Web-based thesaurus which features an associative search for medical researchers. In the course of the thesaurus development, a mouse-over dictionary for English medical terms has been also made available. In addition, we have applied the thesaurus to resolve various drug names appeared in FDA Adverse Events Reporting System. Our thesaurus will pave the way for clinical informatics that enables knowledge finding from medical documents.

Key words

medical thesaurus, Medical Subject Headings, associative search, Adverse Events Reporting System, text mining, mouse-over dictionary