生命科学知識の連想検索における提示語の最適化

金子周司 [1)]，藤田信之 [2)]，鵜川義弘 [3)]
1）京都大学大学院薬学研究科，2）製品評価技術基盤機構バイオテクノロジー本部，
3）宮城教育大学環境教育実践研究センター
email: skaneko@pharm.kyoto-u.ac.jp

連想検索で提示される用語は，テキスト中での特徴的な用語の関連性に基づいて選択されるが，通常の形態素解析では専門用語が過分割され，それらの関連性は専門家にとって周知のものになってしまう。本研究では，専門家が連想検索を使う場合に，より深い専門知識を獲得することができる提示語の選択方法について，階層構造のシソーラスを利用した試みを紹介する。専門用語は，ライフサイエンス辞書に基づくシノニム辞書から一般的な用語を除外した上で，医学論文抄録で最長一致する用語を統制語に置き換えて抽出した。それらの関連性を tf-idf 法で求めた結果，ほぼ妥当な内容が得られたが，一部において連想検索としては価値の低い，シソーラスツリーのパスで近接する用語が散見された。シソーラスを用いて関連語を抽出する場合，用語の類似度を計算式に取り入れる必要性が指摘される。

１．はじめに

ライフサイエンス辞書（LSD, Life Science Dictionary）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]。我々は現在，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，まずは米国の MeSH [2]に準拠する形で始めている。このシソーラス化の目的は大きく 2 つあり，その 1 つは人による情報検索を手助けするような使い道[3]，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[4]である。本研究では，この前者において，検索ワードに対して関連性の高い別のワードを提示する，いわゆる連想検索をシソーラスが存在する状況で最適化しようとした試みを紹介したい。

２．制作前の予備調査

連想検索で提示される用語はコーパスの共起解析によって得られる [5]。そこでまず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，以下の改善点が指摘された。

(1) 過分割による不要語

おそらく汎用辞書を用いた形態素解析で用語を選択していると思われる場合に，例えば「糖尿病」に対して提示される関連ワードに「病」や「性」など，過分割のために生じた不必要な語が多く含まれる場合があった。このことから，解析に用いる辞書には適切な長さの用語を十分に揃えて未知語としての過分割を防ぐ必要性が示唆された。

(2) 関心を呼ばない提示語

Web ページの解析に由来すると思われる場合，例えば「コラーゲン」に対して世間的な関心を呼ぶ関連語ではあるが，科学的には重要視されない「シワ」「美容」のような用語が並ぶ場合があった。これから，解析対象とするコーパスを利用者の関心や用途に応じて選択することが重要であることが示唆された。

(3) 同義語の重複

例えば検索語「モルヒネ」に対して「痛み」と「疼痛」という同義語が並んで提示されるよ

うな例があった。また，対訳関係にある同義語「高血圧」と「hypertension」が重複するような場合も多く見られた。これは対訳シソーラスを解析辞書に用いることで解決でき，その結果，より少ない用語数で広範な関連ワードを提示できると考えられた。しかし，検索語セットを渡す相手にシソーラスを解釈する能力がない場合には，不採用とする用語が含まれる文書が絞り込み結果に含まれない危険があることに留意する必要があるだろう。

3．共起解析

以上の調査から，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。

そこで，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，解析辞書としては日本語および英語の計 20 万種類の表記を 27,562 語の統制語（英和対訳）に集約した 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（図 1）。

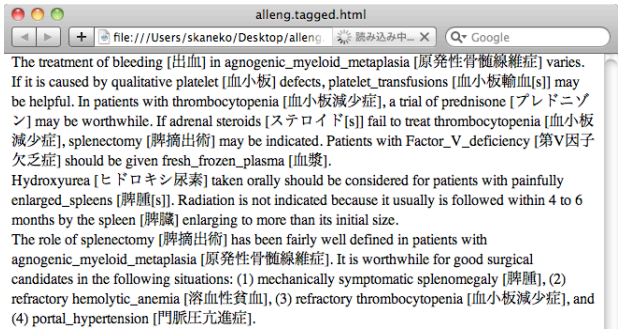

図 1　タグ付けの例

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-idf [9]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 (83%)語，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 50 語を各統制語について付与する関連ワードとした（図 2）。

片頭痛

Migraine Disorder　　総共起数　849

頭痛	97	241
疼痛	75	134
前兆を伴う片頭痛	34	123
てんかん	52	121
スマトリプタン	32	115
群発頭痛	27	103
症状	58	79
トリプタミン	19	67
カルシトニン遺伝子関連ペプチド	24	65
皮質拡延性抑制	17	63
エルゴタミン	13	54
悪心	22	54
セロトニン	25	48
ゾルミトリプタン	10	43
ジヒドロエルゴタミン	8	35
メチセルジド	9	35
神経学	13	35
トピラマート	9	33
エストロゲン	17	32
経口避妊薬	11	31

図 2　「片頭痛」と共起する統制語
（数値は左が共起頻度，右が tf-idf 値）

なお，臨床症例報告のみを収集した同サイズのコーパスを用いて同様の解析を行った結果，診断および治療法を表す用語と治療薬名が非常に多く検出された。LSD ユーザーとして臨床家より基礎研究者が多いという実情を踏まえ，公開辞書に実装するデータとしては先に行った代表学術誌の分析結果に基づくこととした。

4．実装

図3 WebLSD のシソーラス表示

従来より公開している WebLSD [10]の英和および和英辞書から「シソーラス」のリンクを設けた。シソーラスは別ウィンドウにて図 3 のような情報とともに関連語を提示した。

「LSD シソーラス:」表示に続いて統制語を和英表記し，続いて同義語（異表記），概念ツリー（規定済みの場合），最後に関連ワードを連想検索として表示している。医薬品の場合には化学構造や薬理作用情報も併せて表示される。

連想検索での共起語からは，日本語から Google へのリンク，英語から NCBI Entrez へのリンクを現在は設けている。

５．考察

連想検索として提示した関連語の内容について，図 3 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に思われる。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，図 3 の例においては中位以下に「エストロゲン受容体」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例えば図 2 で見られるような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，次に述べる点でさらに改善の余地があることが指摘できる。

(1)概念ツリーで近接する用語は不要である

シソーラスの概念ツリーを明示している場合に，連想検索において上位語（図 3 の例では「睡眠異常」）を提示しても用語の定義以上の意味はもたらされない。他にも医薬品や生体分子の関連語として，類似薬や薬品分類などの上位語あるいは同位語が含まれていた例が多く，これらは連想の範囲を狭める直接的な関連語と言える。概念ツリーに明示される用語は類似度計算などを用いて排除することが望ましいと考えられる。

(2)検索がヒットしない用語の混入がある

列挙した関連語の中に，リンク先において関連文書や論文がヒットしない例があった（図 3 においては「Gender Identity」「Propaganda」などが該当する）。PubMed の解析結果に基づいて選択した関連語から PubMed リンクが得られない原因の多くはリンク時に統制語を用いたことに起因すると考えられるため，統制語の選び方やリンクの渡し方を工夫する必要があるだろう。また，一部の用語については略語を解析辞書に採用した結果，異なる意味の同表記を抽出した失敗も含まれていたようである。

本研究では連想検索における関連語の最適化を図り，公開辞書サービスへの実装を試みた。2009 年 4 月から 9 ヶ月間の統計でみると，オンライン辞書の総検索 3,500 万件中の 125 万件（3.6%）でシソーラスが参照されている程度であり，連想検索の利用度は不明であった。しかし用語の訳を調べるだけでなく，情報ポータルとしての辞書利用は着実に増えつつあり，一定の役割を果たしうるものと期待できる。今後もシソーラスの充実を図るとともに，より有用な連想検索を実現できるように検討を続けていきたい。

参考資料

[1] 金子周司，ライフサイエンス辞書とは．情報管理 49 (1): 24-35, 2006.
[2] http://www.ncbi.nlm.nih.gov/mesh
[3] 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効率的医療情報検索システムの開発．医療情報学，28 (Suppl.): 639-642, 2008.
[4] 金子周司，薬物有害事象 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，46 (1): 125-132, 2010.
[5] 長尾真編，岩波講座ソフトウェア科学 15 自然言語処理, pp.387-391, 1996.
[6] Webcat Plus (http://webcatplus.nii.ac.jp)
[7] reflexa (http://labs.preferred.jp/reflexa)
[8] PubMed (http://www.ncbi.nlm.nih.gov)
[9] 長尾真編，岩波講座ソフトウェア科学 15 自然言語処理, pp.419-421, 1996.
[10] http://lsd-project.jp/