医薬品安全性に関する文献情報自動抽出システムの考案

天野 博夫　金子 周司
京都大学大学院薬学研究科生体機能解析学分野

A newly devised text search system for adverse drug reactions.

Amano Hiro Kaneko Shuji
Department of Molecular Pharmacology, Graduate School of Pharmaceutical Sciences, Kyoto University.

In recent years, drug safety has become a major issue for those engaged in medical care. Although medical literatures are on a watch list for drug safty matter, it is a tremendous task to sort enormous amount of text information. We devised an exhaustive text search system for specialized use in pharmacovigilance termed TSADR (Text-search System for Adverse Drug Reaction), which is made of two medical vocabularies (DN-INDI list and RN list) and a Perl script file. TSADR extracts sentences involving topics relevant to the drug safety from PubMed abstracts, creates HTML files which show extracted texts with drug names and adverse reaction names color-coded on pages in a web browser and assist searchers to discriminate important items. TSADR is now developing toward the practical use for text analysis in pharmacovigilance to identify and anticipate adverse reactions resulting from drug use.

Keywords: adverse drug reaction, literature information

1. 研究の背景と目的

製薬企業のグローバル化が進行し、新薬の開発・供給体制の迅速化が図られる一方で、医療関係者共通の重要問題である薬の副作用等、安全性に関する情報の収集・伝達体制の整備は必ずしも進んでいない。文献情報の収集・解析により副作用の発生を早期に検知あるいは予測するためには、質的なばらつきが大きい大量のテキストソースを網羅的に検索して医薬品の安全性に特化した情報を選別・抽出する作業が必要であり、これは通常のキーワード検索では事実上不可能である。本研究においては、PubMedアブストラクトを対象として、医薬品の安全性に関する情報の選別・抽出を補助するシステムの構築を目的とした。

2. 研究方法

米国FDAからAERS(Adverse Event Reporting System)データベース[1]2004年第一四半期から2005年第二四半期まで1年半分のASCIIデータファイルをダウンロードし、"DRUGNAME"、"INDI-PT"、"PT"各フィールドのデータから医薬品名とその適応を関連させたリスト(DN-INDI list)および有害反応名のリスト(RN list)を作成した。これらを辞書としてPubMedアブストラクトから医薬品名と有害反応名が同時に記載されているセンテンスを抜き出し、ヒットした用語の認識性を色別表示により改善するPerlスクリプト(TSADR)を作成した。適応名の記載が同一のセンテンス内にあれば、これも表示させた。TSADRの基本的な動作の概要をフローチャート(図1)に示す。PubMedアブストラクト はLimits設定フォームにおいてonly items with abstracts, English, Humansの3つの制限のみを設定し、キーワードを入力せずに取得した500件分のテキストを検索対象の1単位とした。TSADRにより抽出されたセンテンスを含むアブストラクトを読み、医薬品の副作用の記載が正しく抽出・表示されているかを検証した。1単位のテキストサンプルに対する一回のオペレーションの抽出率(500件中何件のアブストラクトが抽出されたか)および正解率(抽出されたアブストラクトの何%に医薬品の安全性に関する内容が記載されていたか)をシステムの評価基準とした。

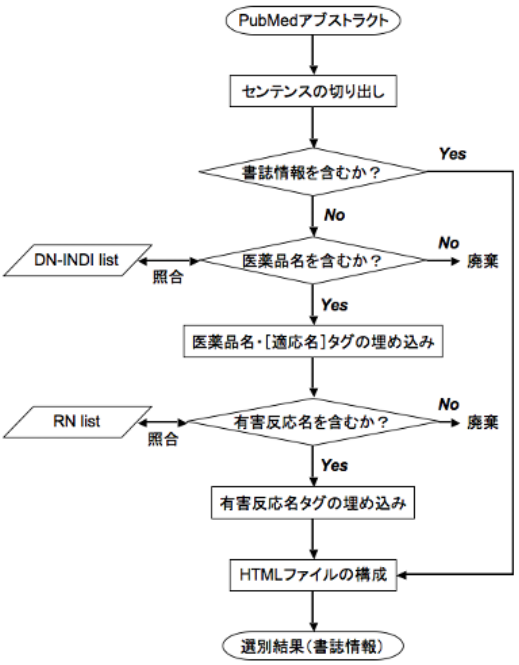

図1 TSADR基本動作のフローチャート

3. 研究成績と考察

図2は、上記研究方法に取得条件を示した、ヒトに関する英文アブストラクト付きPubMedアイテムの年間エントリー数の推移を示したものである。年を追ってエントリー数の増加が認められるが、2005年のデータを参考にすると、このコーパスから網羅性を重視して

必要なアイテムを選別していくには、一日平均800ないし1000件を処理する必要があり、その実行には豊富かつ高水準の労働力、または自動化による補助が必須であると考えられる。

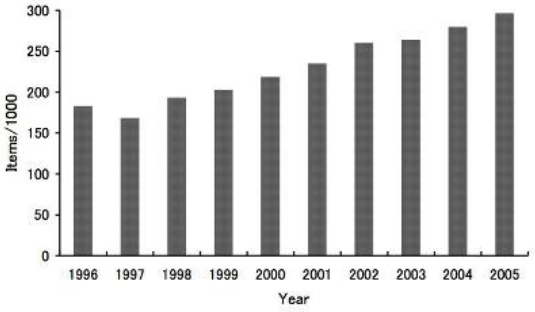

図2 PubMedアイテムの年間エントリー数

研究方法に記載の取得条件に合致する、ヒト関連の英文アブストラクト付きPubMedアイテム年間エントリー数の推移

本研究において医薬品の安全性に関連する語彙のソースとして用いたAERSデータベースは、米国FDAが収集・管理している、製薬企業からの義務報告と医療従事者・患者およびその家族からの自発報告を総合した巨大な副作用データベースであり、四半期分(収載報告件数8-9万)ごとに半年遅れの生データがASCIIまたはSGMLファイルとしてFDAのサイトから入手できる。今回は2004年および2005年上半期の一年半分のレコードをベース語彙リストのソースとして用いた。AERSデータベースはリレーショナルデータベースの構造を持つため、医薬品名と適応名を対応させて取得できる大きなメリットがある。初期システムにおいては、3813の医薬品名(drug name)、3824種類の適応(indication)をDN-INDI listに、9712種類の有害反応名(reaction name)をRN listに収載した。

初期システムを用いて、日本時間06年5月29日に取得したPubMedアブストラクト500件(11,924センテンス)より74件(138センテンス)が抽出された(抽出率14.8%)。74件中、有害反応名が正しく表示されていたものが26件(正解率35.1%)、32件では医薬品名と直接には無関係な反応がヒットし、16件は適応症が有害反応として誤って表示されていた。誤った選別のパターンとしては、"glucose"、"oxygen"、金属イオン等の生体成分が医薬品名として拾われて起こる事例や"alcohol"、"antibiotic"、"chemotherapy"等、医薬品分類名に関して誤った選別が起こる例が多く認められた。前者のパターンに関しては、隣接する単語との関連から医薬品名としての取捨を判断するフィルターをスクリプトに加えて対応し、後者のパターンはDN-INDI listの適応エントリー数を増やす手段で対応した。システムの構造上、語彙リストの医薬品名、有害反応名のレコード数を増やせば抽出率は上昇し、検索の網羅性に関しては有利に働く一方で、副作用以外の医薬品名と有害反応名の組み合わせを拾う可能性も高くなり、正解率が下がれば人間による最終的な選別操作の負担が大きくなる。これに対して、医薬品を投与する原因となる病態などの名称、すなわち適応名の語彙を増やすことは、誤った選別を抑制し、検索の精度を上昇させる。

上記対応を施したシステムを、新たに(日本時間06年6月14日)取得したシステムトレーニング用テキスト(STTXT)に適用し、その結果を基に語彙リストファイルを修正する作業を繰り返した。また、誤って選別されたアブストラクトに癌・腫瘍関係の雑誌のものが多かったことから、癌・腫瘍関係語彙用テキスト(CTBTXT: PubMedからSubsetsのLimitにCancerを設定して取得した)を用いたトレーニングも行った。これらのトレーニングによって最適化された語彙リストを用いて、両トレーニングテキスト自身を検索した最終的な成績(正解件数／抽出件数)はSTTXTが35/62(正解率56.5%)、CTBTXTが42/64(正解率65.6%)であった。必要な場合にはスクリプトの修正も行った。

本システムの実用化に向けて、システムパフォーマンスに対する上記トレーニングの有効性を検討する目的で、新たに取得したテキスト(日本時間06年8月1日)をトレーニング前のシステムTSADR-originalとトレーニング後のシステムTSADR-trainedを用いて解析し、得られた成績を比較した。TSADR-original、TSADR-trainedそれぞれの成績(正解件数／抽出件数)は14/44(正解率31.8%)および22/54(正解率40.7%)と算出され、抽出率、正解率ともにトレーニングの有効性が認められた。

一方、選別されるべきセンテンスの拾い漏れがどの程度起こっているかの予備的検討として、医薬品文献情報の有力サイトである英国のNational electronic Library for Medicines[3]に最近(Date Published 12/04/2006-24/08/2006)ピックアップされた項目のうち副作用情報に分類される120レコードを対象にTSADRによる重要文献の抽出漏れを検討した。120件中抽出されなかったレコードは14件であった(抽出率88.3%)。抽出漏れの原因としては、DN-INDI listに医薬品名が収載されていなかったものが11件、RN listに有害反応名が収載されていなかったものが5件、2件は医薬品名、有害反応名ともリストに記載はあったが、別個のセンテンス中に記載されていたためヒットしなかった。

医薬品の安全性に関する情報は、新薬に関してその重要性が特に大きく、本検索システムの実用性は医薬品名を主とする語彙リストのアップデート状況に強く依存する。今後、本システムの網羅性および選別性をさらに向上させるために、語彙リストの補強・改訂を自動化する方法を検討する予定である。

参考文献

[1] Adverse Event Reporting System(AERS).http://www.fda.gov/cder/aers/default.htm.
[2] Entrez PubMed.http://www.ncbi.nlm.nih.gov/entrez/query.fcgi.
[3] National electronic Library for Medicines.http://www.druginfozone.nhs.uk/home/default.aspx.