# 自然言語処理技術による副作用関係抽出
# Extraction of Adverse Effect relation by using natural language processing

## 要　旨

臨床試験を経て認可を受けた医薬品は、製造販売後の使用成績調査が義務づけられている。このような調査はほとんどの場合、医薬品を販売する製薬会社が行うが、医薬品を投与し、患者の様子を報告するのは病院で診療を行う医師である。従って、この市販後調査には製薬会社に大きなコストが発生する。とりわけ、医薬品の副作用に関する項目は、医師の記憶や診療録などを参照して埋める必要があるため、診療行為で多忙な医師にも重い負担がかかる。

本研究では、用語抽出、関係抽出、表記ゆれ解消、辞書拡張など複数の異なる言語処理の要素技術を統合して、医薬品における副作用出現状況の調査を支援するシステムを構築した。このシステムは、退院時サマリーから、副作用に関して記述されている箇所を特定し、さらに、医薬品や副作用症状ごとに集計する機能を備えている。

本稿ではシステム全体の構成と個々の要素技術について解説する。

## Abstract

Post Marketing Surveillance (PMS) is required for approved drugs that have undergone clinical trials. Pharmaceutical companies mainly conduct surveillance, while doctors administer drugs and report the condition of patients. PMS imposes a tremendous cost on pharmaceutical companies. It also imposes a heavy burden on doctors who are busy attending on patients, as details of ADRs need to be described by relying on a doctor's memory or referring to clinical records.

In view of this background and to support the survey of ADRs, we have developed a system to collect ADRE data from discharge summaries in Japanese created in a hospital.

執筆者

大熊　智子（Tomoko Ohkuma）
三浦　康秀（Yasuhide Miura）
外池　昌嗣（Masatsugu Tonoike）
杉原　大悟（Daigo Sugihara）
増市　博（Hiroshi Masuichi）

研究技術開発本部 コミュニケーション技術研究所
（Communication Technology Laboratory , Research & Technology Group）

# 1. はじめに

## 1.1 背景

臨床試験を経て認可を受けた医薬品は、製造販売後の使用成績調査が義務づけられている。このような調査はほとんどの場合、医薬品を販売する製薬会社が行うが、医薬品を投与し、患者の様子を報告するのは病院で診療を行う医師である。従って、この市販後調査には製薬会社に大きなコストが発生する。とりわけ、医薬品の副作用に関する項目は、医師の記憶や診療録などを参照して埋める必要があるため、診療行為で多忙な医師にも重い負担がかかる。

## 1.2 本研究の目的

本研究では、このような背景のもと、医薬品における副作用出現状況の調査を支援するために、複数の言語処理技術を統合してシステムを構築した。このシステムは、退院時サマリーから、副作用に関して記述されている箇所を特定し、さらに、医薬品や副作用症状ごとに集計する機能を備えている。本システムは用語抽出、関係抽出、表記ゆれ解消、辞書拡張など複数の異なる言語処理の要素技術を組み合わせて実現される。

本稿の構成は以下のとおりである。2 章ではシステム全体の構成について説明する。3 章では副作用関係抽出について述べる。4 章では副作用関係が認められた医薬品と副作用表現の標準化について述べる。5 章では標準化された副作用表現と、薬効に対応づけられた医薬品を直行表で表示する機能について述べる。6 章では今後の課題や活動計画について述べる。

# 2. システムの構成

図 1 に副作用関係集計システムの構成図を示す。このシステムは大きく 3 つの機能から構成される。

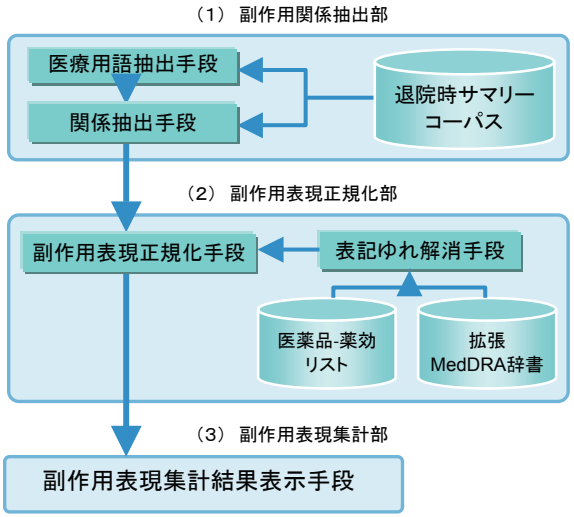


図 1. システムの構成
Overview of Adverse-drug event extraction system

（1）副作用関係抽出部
このシステムに退院時サマリーのテキストデータが入力されると、テキスト中に含まれる医薬品名、疾患名、検査名などの各種医療表現の抽出が行われる。次に抽出された医療表現を入力として副作用関係が成立する医薬品名と疾患名のペアを Support Vector Machine（SVM）で判定する。

（2）副作用表現正規化部
副作用関係が成立したペアに含まれる副作用表現と医薬品名の正規化を行う。医薬品名に対しては、人手で薬効分類コードを付与した。副作用表現に対しては辞書と表記ゆれ解消手段を用いて、副作用記述のための用語集である MedDRA/J[*1] に対応づけた。

（3）副作用表現集計部
正規化された副作用表現と医薬品名を集計し、直交表として表示する。

次の章からは、これらの 3 つの機能についてより詳しく説明する。

[*1] MedDRA/J は、Medical Dictionary for Regulatory Activities/Japanese version　ICH 国際医薬用語集日本語版のことである。4.2 にて詳細説明。

## 3. 副作用関係抽出

### 3.1 機械学習のための退院時サマリーコーパス

退院時サマリーとは患者の入院以前の経緯や入院中の経過が簡潔に記された文章である。多くの場合、時間軸に沿って、患者の状態、処置した内容、検査結果等のうち、臨床的に重要だと判断されたものが記載される。本研究では、退院時サマリー464 件のテキストに対し、医療表現やモダリティー副作用関係情報のアノテーションを行い、機械学習に用いている。

### 3.2 医療表現抽出

入力された退院時サマリテキストを、文字単位の IOB2 形式[*2] に変換し Conditional Random Fields（CRF）による機械学習を用いてタグ系列を推定した。素性としては、形態素解析結果など標準的な素性を用いた。ここで抽出した医療表現の種類は 13 種類である。そのうち、医薬品名と副作用表現（疾患名、変化、検査結果）の精度を表１に挙げる。

表 1. 医薬品名、副作用表現の抽出精度
Accuracy of Drug name/Symptom name extraction

| 医療表現 | 適合率(%) | 再現率(%) | F 値 |
|---|---|---|---|
| 医薬品名 | 86.9 | 81.3 | 84 |
| 疾患名 | 85.5 | 80.2 | 82.8 |
| 変化名 | 84.6 | 74.8 | 79.4 |
| 検査結果 | 80.7 | 76.3 | 78.4 |

### 3.3 副作用関係抽出

本稿では副作用関係を「医薬品名 D と症状名 S の間に副作用の関係が認められる」関係と定義する。図２に副作用関係の例を示す。

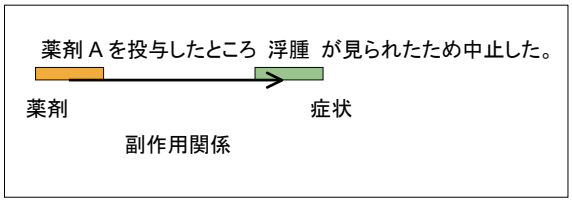


図 2. 副作用関係抽出の例
Example of Adverse Drug Event

副作用関係抽出は SVM による機械学習[2)]によって行った。学習データは一文中に医薬品と副作用表現が含まれる文を対象にした。１）では、文中に出現する医薬品と疾患名、変化名、検査結果のペアのうち、副作用関係が成立するペアを正例、副作用表現ではないもの、そして副作用表現ではあるものの否定表現であるものを負例として学習する。

副作用関係ペアを特徴付ける素性としては、表２の素性を用いる。なお、ペア間係り受け最短パスとは、医薬品を含むチャンク（句、文節）と副作用症状候補を含むチャンクの係り先をたどった際に、同じチャンクに至るまでのパスを意味する。副作用関係抽出の識別精度を表３に示す。

表 2. 副作用関係抽出のための素性
SVM features for Adverse-drug event extraction

| 素性 | 説明 |
|---|---|
| 文字距離 | 医薬品と副作用症状候補間の文字数 |
| 形態素距離 | 医薬品と副作用症状候補間の形態素数 |
| 出現順序 | 医薬品の後に副作用症状が現れる場合は真、逆であれば偽 |
| ペア間形態素 | 医薬品と副作用候補の間に現れる形態素の原形 |
| ペア間係り受け | 医薬品と副作用症状 |
| 最短パス候補の係り受け | 解析結果での最短パスに含まれるチャンク中の形態素の原形 |
| 文中の医療表現 | 副作用関係ペアが現れる文中に存在する医療表現 |
| 副作用症状候補中の医療表現 | 副作用症状候補と入れ子関係にある場合の医療表現 |

表 3. 副作用関係抽出の精度
Accuracy of Adverse-drug event extraction

| 抽出対象 | 適合率(%) | 再現率(%) | F 値 |
|---|---|---|---|
| 疾患名のみ | 39.8 | 56.4 | 46.7 |
| 上記以外 | 28.5 | 23.9 | 26 |

[2] IOB2 形式とは、あるひと固まりの文字列（チャンク）にタグを付与するための形式である。B はチャンクの開始位置（Begin）、I はチャンクの一部（Inside）、O はチャンクに含まれない（Outside）を意味する。本稿では医療用語をチャンクの単位としている。

## 4. 医薬品名と副作用表現の標準化

### 4.1 医薬品名に対する薬効分類コード付与

副作用を起こしている原因として抽出された医薬品名に対し医薬品の使用目的の分類体系である 3 桁の薬効分類コードを人手で付与した。なお、１つの薬品が複数の薬効分類コードを持つ場合もあるため、医薬品と薬効分類コードの対応はいつも一対一であるとは限らない。

### 4.2 MedDRA/J について

MedDRA とは日米欧医薬品規制ハーモナイゼーション国際会議（ICH）によって開発された階層構造を持つ医学用語集であり、MedDRA/J はこの日本語版である。MedDRA/J は SOC（器官別大分類）、HLGT（高位グループ用語）、HLT（高位用語）、PT（基本語）および LLT（下層語）の５階層構造となっている。国内では、医薬品の副作用用語として MedDRA/J を使用することが推奨されている。例えば薬剤市販後調査データベースにおいても、調査票の集計に MedDRA/J を用いている[4)]。

しかしながら、MedDRA/J を今回我々が目的とする副作用関係の抽出にそのまま利用するのには２つの問題があった。

まず１つは、MedDRA/J に収載されている語に多様性があることである。MedDRA/J に収載されている語は副作用症状について記載するための語彙であり、その収載語が必ずしも副作用症状を表すものであるとは限らない。例えば、「離婚した両親」、「教育問題」など社会的背景に関する語や「身長」、「体重」など身体の測定に関する語など、登録されている語はさまざまである。従って、MedDRA/J 収載語のすべてを副作用症状を表す用語として用いることができない。

もう１つの問題は、MedDRA/J のカバーしている語彙と診療録で実際に用いられる用語に相違があることである。上述のように、副作用調査の報告に MedDRA/J の使用が推奨されていても、診療録などの調査の元になるデータが MedDRA/J を意識して記述されているわけではない。従って、副作用に関する記述が診療録中のテキストにあったとしても、それが MedDRA/J に収載されている語彙である保証はない。これらの問題を解決するために、MedDRA/J の精査と拡張を行った。

### 4.3 MedDRA/J の精査

MedDRA/J に登録されている用語を疾患とそれ以外の語に分類する作業を行った。最初のステップとして、MedDRA/J の最上位階層である SOC のうち、疾患以外の語である可能性が高く、除外しても問題ないものを調査した。その結果、２つの SOC 分類に含まれる LLT3,860 語は疾患名ではないことが分かったが、それ以外の語は疾患名や副作用表現を多く含むため、SOC による分類ができないことが分かった。SOC の次の階層構造である HLGT はほぼ疾患名であり、その次の階層である HLT は 1,699 項目あるため、精査の手がかりにするには適さないと判断した。

次に、電子カルテ用の標準病名用語集（標準病名マスター）と LLT を比較し、標準病名マスターに含まれる LLT4,552 語を疾患名として抽出した。さらに、ヒューリスティクスに基づくルールを用いて、語の分類を行った。その結果、LLT6,363 語を疾患名として抽出することができた。さらに自動分類によって分類ができなかった語に対しては人手で分類を行った。

以上の精査の結果、MedDRA/J に登録されている語彙は表 4 のように分類された。

表 4. MedDRA/J の分類結果
Classification of terms in MedDRA/J

| 分類 | 語数 | 精査手法 | 語数 |
|---|---|---|---|
| 疾患を表わす表現 | 13,815 | 病名マスターと一致 | 4,552 |
| | | ルールによる分類 | 6,363 |
| | | 人手による分類 | 2,900 |
| 疾患以外の表現 | 4,638 | MedDRA の階層を利用した分類 | 2,087 |
| | | ルールによる分類 | 1,355 |
| | | 人手による分類 | 1,196 |
| 不明 | 30 | N/A | N/A |

### 4.4　MedDRA/J の拡張

診療録のテキストには副作用症状を表現する際に MedDRA/J には登録されていない語が用いられることもある。そこで、副作用を表現する疾患を以下の手法で MedDRA/J に新たに追加した[5]。

まず、MedDRA/J に追加する語彙の候補として、薬剤添付文書の副作用欄に列挙されている語を抽出した。次に、表記ゆれ解消手法を用いて、抽出した語彙候補と MedDRA/J の PT の対応づけを行った。対応づけに対して人手でチェックを行った結果、自動的に出力された結果 580 語のうち、437 語が正しく対応づけられていることが分かった。

上記の自動手法で LLT との対応づけができなかった語に対して、最大 50 語の対応先 LLT 候補を出力した。その結果、これらの候補から MedDRA/J への正しい対応先が得られた語は 489 語だった。表 5 に追加した語の例を示す。

表 5.　MedDRA/J に新規に対応づけられた語の例
Example of new term corresponding to LLT

| 元の語 | LLT | PT |
|---|---|---|
| ビリルビン上昇 | ビリルビン値上昇 | 血中ビリルビン増加 |
| BUN 上昇 | BUN 増加 | 血中尿素増加 |
| 身ぶるい | 身震い | 振戦 |
| 血中 Ca 減少 | 血中カルシウム減少 | 血中カルシウム減少 |

## 5.　副作用表現の集計

副作用表現集計部では、MedDRA/J 収載用語に対応づけられた副作用表現と薬効コードを付与された医薬品名を集計し、直交表として表示する。図 3 に直交表の例を示す。なお、本稿では個人情報保護の観点から、テキストが判読できない画像を使用している。調査を実施するユーザーは、集計件数を参照するだけではなく、個別の薬効や副作用を手がかりにして、より詳しく調べたい退院時サマリーのテキストに直接アクセスすることができる。

## 6.　今後の課題

### 6.1　同一文中に現れない副作用関係の抽出

3章の副作用関係抽出技術では、関係抽出の対象を同一文内に医薬品と副作用表現が出現するものに限定している。しかし、実際にはこの条件に該当するものは全体の約 20%ほどである。今後は、複数文にまたがる副作用関係[6]も抽出対象にしたい。

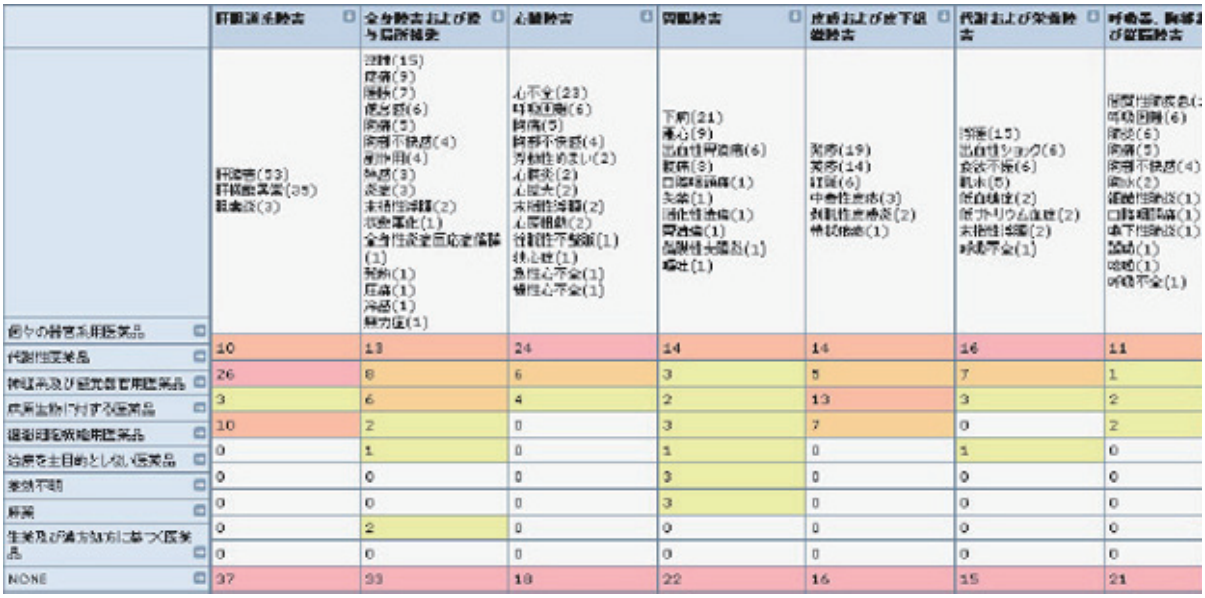


図 3.　直交表の例
Example of the Cross Table

### 6.2 多様な副作用表現の正規化

本稿では、副作用表現を MedDRA/J 記載用語に対応づけるために、表記ゆれ解消技術と辞書の拡張という２つの手法を用いた。しかし、これらの技術が適応できる範囲は、副作用表現が体言であることが前提となる。例えば「足がずきずきと痛い」という表現を「下肢疼痛」という用語に対応づけることは難しい。また、「心カテ」と「心臓カテーテル」、“WBC” と “White Blood Cell” など略語として現れる表現も現在は扱えていない。今後は、7）や 8）などの略語展開技術の適用や多様な副作用表現の正規化のための副作用表現コーパス構築を検討していきたい。

### 6.3 病院の違いによる差異の分析

今回の実験では、機械学習用のデータと副作用判定は同じ病院の退院時サマリーを用いた。しかし、機械学習を行った病院とは別の病院のデータで本稿の処理を行ったときに、差異がどのくらいあるのかを検討する必要がある。今後は、別の病院の退院時サマリーを対象にした実験を行い、病院間の共通点と差異について分析したい。

本研究では、医薬品の副作用の調査を支援するための統合的な言語処理システムを実現した。このシステムは退院時サマリーから、副作用に関して記述されている箇所を特定し、さらに、医薬品や副作用症状ごとに集計する機能を備えている。ユーザーは副作用表現と医薬品の直交表という UI によって、調査したい事例に直接アクセスすることが可能になる。

今後は、今回抽出対象外とした同一文中にない副作用関係の抽出、多様な副作用表現の正規化、そして他病院での実験などに取り組む予定である。

本研究は東京大学附属病院大江和彦教授と東京大学知の構造化センター荒牧英治特任講師との共同研究の成果である。

## 7. 参考文献

1） Yasuhide Miura, Aramaki Eiji, Tomoko Ohkuma, Masatsugu Tonoike, Hiroshi Masuichi, and Kazuhiko Ohe. Adverse-effect relations extraction from massive clinical records. In COLING 2010 Workshop (In cooperation with Infoplosion) The Second International Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010), 2010
2） くすりの適正使用協議会薬剤疫学部会 PE 研究会。経口抗菌剤の使用成績調査データベースの構築 ―最終報告―、2007
3） 三浦康秀、荒牧英治、大熊智子、外池昌嗣、増市博、大江和彦。複数文にまたがる関係抽出における構文情報の効果。言語処理学会第 17 回年次大会、2011
4） 山田恵美子、荒牧英治、外池昌嗣、大熊智子、三浦康秀、杉原大悟、増市博、大江和彦。文脈情報を用いた略語の曖昧性解消。第 30 回医療情報学連合大会、2010
5） 篠原（山田）恵美子、三浦康秀、外池昌嗣、大熊智子、増市博、荒牧英治、大江和彦。共起・連接頻度グラフに基づいた略語展開語候補生成。言語処理学会第 17 回年次大会、2011
6） 杉原大悟、大熊智子、三浦康秀、外池昌嗣、増市博、山田恵美子、荒牧英治、大江和彦。表記ゆれ解消手法を利用した副作用表現の獲得。第 30 回医療情報学連合大会、2010
7） 山田恵美子、荒牧英治、外池昌嗣、大熊智子、三浦康秀、杉原大悟、増市博、大江和彦。文脈情報を用いた略語の曖昧性解消。第 30 回医療情報学連合大会、2010
8） 篠原（山田）恵美子、三浦康秀、外池昌嗣、大熊智子、増市博、荒牧英治、大江和彦。共起・連接頻度グラフに基づいた略語展開語候補生成。言語処理学会、第 17 回年次大会、2011

筆者紹介

大熊　智子
研究技術開発本部コミュニケーション技術研究所に所属
専門分野：自然言語処理　国語学

三浦　康秀
研究技術開発本部コミュニケーション技術研究所に所属
専門分野：自然言語処理

外池　昌嗣
研究技術開発本部コミュニケーション技術研究所に所属
専門分野：自然言語処理　音声処理

杉原　大悟
研究技術開発本部コミュニケーション技術研究所に所属
専門分野：自然言語処理

増市　博
研究技術開発本部コミュニケーション技術研究所に所属
専門分野：自然言語処理