電子カルテからの副作用関係の自動抽出

三浦　康秀[†]　荒牧　英治[‡]　大熊　智子[†]
外池　昌嗣[†]　杉原　大悟[†]
増　市　博[†]　大江　和彦[††]

[†]富士ゼロックス株式会社 研究技術開発本部　[‡]東京大学 知の構造化センター
[††]東京大学 医学部附属病院
yasuhide.miura@fujixerox.co.jp, eiji.aramaki@gmail.com, {ohkuma.tomoko, masatsugu.tonoike, daigo.sugihara, hiroshi.masuichi}@fujixerox.co.jp, kohe@hcc.h.u-tokyo.ac.jp

1. はじめに

近年，電子カルテの普及に伴って大量のカルテ文書が電子データとして蓄積されるようになった．これらは患者の健康についての重要な臨床情報を含み，診療・研究のための情報抽出が望まれている．しかし，臨床情報はその多様性ゆえ自然言語で記述されることも多く，その抽出は容易ではない．

電子カルテに自然言語で記述される重要な臨床情報の 1 つとしては，患者への投薬情報およびそれに関連する副作用情報が挙げられる．医薬品の副作用は臨床試験により事前に調査されているが，医療現場で実際に発生する副作用は患者の状態および同時に服用される別の医薬品の影響を受ける．このため，医薬品の投与により症状が現れても，どの医薬品の副作用であるか，またそもそも副作用であるかは必ずしも明確ではない．このような実際に発生した副作用情報は，電子カルテの退院時サマリの入院後経過欄には治療経過と共に自由記述されている．

本研究では，この副作用情報を医療テキストから機械学習手法を用いて自動的に抽出することを目指した．医薬品の副作用を “医薬品” と “症状” 間の関係 (relation) として捉え，MUC[1)]，ACE[2)]，SemEval[3)]，BioCreative[4)] で扱われている，関係識別・関係抽出問題として解くことを考えた．本稿では，副作用関係を抽出する対象のテキストおよびそのアノテーションの枠組み，副作用関係抽出手法，評価実験結果について述べる．

2. 副作用関係を抽出する対象のテキスト

副作用関係を抽出する医療テキストとしては，2,577 件の退院時サマリの入院後経過セクションの自由記述欄を用いた．退院時サマリとは患者の退院時に医療従事者により記述される患者の記録であり，入院後経過セクションには患者の病院での治療経過が記述されている．この 2,577 件の退院時サマリは，A 病院の全診療科から集められた退院時サマリから HIPAA のガイドライン☆に基づき個人情報を削除し，キーワード “中止”，“変更”，“副作用” を含むものをランダムに選別した．キーワードを指定

☆ http://www.hhs.gov/ocr/privacy/

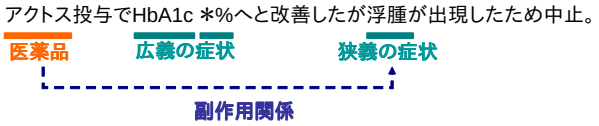

図 1　副作用関係の例

表 1　副作用症状候補タグの定義・例

種類	定義・例
狭義の症状	具体的な疾患名・症状名として副作用を直接的に表す．例：呼吸困難感，浮腫，発熱
広義の症状	検査値・状態変化表現として副作用を間接的に表す．例：1.05mg/dl(検査値)，HbA1c，悪化，体重増加

しているのは，人手によるアノテーションのコストを削減するためである．退院時サマリには副作用の記述がないものが多く，現在は効率的にアノテーションを行うためにキーワードを含むサマリのみを対象にしている．選別された退院時サマリの入院後経過セクションには，4 人のアノテータにより 2.1 節の枠組みの副作用関係情報のアノテーションを行い，副作用関係コーパスを構築した．

2.1 副作用関係コーパス

副作用関係コーパスは，医療表現のアノテーションおよびアノテートされた医療表現間の関係により構成されている．これは副作用関係は，図 1☆☆のように医療表現間の関係として定義できると考えているためである．

医療表現のアノテーション

副作用関係コーパスには 15 種類の医療表現タグをアノテートした．15 中の 14 のタグは，我々が以前行った退院時サマリの現病歴セクションの可視化 [5)] でアノテートしたタグと同じ仕様を用いた．新規のタグとしては，“副作用症状候補” タグを副作用関係を表現するために追加した．このタグは症状を広く表すタグであり，表 1 の “狭義の症状” と “広義の症状” に分けられる．

医療表現間の関係のアノテーション

副作用関係コーパスでは医薬品と副作用症状候補の関係を定義することにより，副作用関係を表現している．タグに relation 属性を付与し，関係を持つタグ同士に同じ

☆☆ 本稿の例文では，個人情報保護の観点から全ての数値表現を “＊” に置き換えている．

<drug relation="1">アクトス</drug>投与で<symptom sense="wide">HbA1c</symptom> <symptom sense="wide">*%</symptom>へと改善したが<symptom relation="1">浮腫</symptom>が出現したため中止。

図 2　副作用関係のアノテーション例

表 2　アノテーション数

アノテーション		数
医薬品タグ		2,739
副作用症状候補タグ	狭義	5,588
	広義	8,681
副作用関係		738

<drug relation="1">アクトス</drug>投与で<symptom sense="wide">HbA1c</symptom> <symptom sense="wide">*%</symptom>へと改善したが<symptom relation="1">浮腫</symptom>が出現したため中止。

副作用関係	医薬品	副作用症状候補
×	アクトス	HbA1c
×	アクトス	*%
○	アクトス	浮腫

図 3　副作用関係ペアの抽出

ID を持たせている．図 1 にアノテーションを付与した例を図 2 に示す．drug が医薬品に対応し，symptom が副作用症状候補に対応し，symptom の sense 属性により狭義 (sense なし) および広義 (sense=“wide”) の症状を表現している．

現在までに副作用関係コーパスにアノテートされた医薬品タグ，副作用症状候補タグ，副作用関係の数を表 2 に示す．

3. 副作用関係抽出手法

副作用関係コーパスでは，副作用関係を医薬品タグと副作用症状候補タグ間の関係として定義している．そこで提案手法では，医薬品と副作用症状候補のペアをコーパスより抽出し，副作用関係にある場合を正例，ない場合を負例として，副作用関係を機械学習手法を用いて学習する．以下，その手続きを示す．

ステップ 1: 副作用関係ペアの抽出

副作用関係コーパスより，同一文中に出現する医薬品と副作用症状候補のペアを，副作用関係ペアとして抽出する．このとき relation 属性の ID が一致するペアを正例，一致しないペアを負例にする．図 3 に正例・負例の抽出例を示す．

ステップ 2: 副作用関係ペアからの素性の抽出

副作用関係ペアを特徴付ける素性としては，表 3 の素性を用いる．なお，ペア間係り受け最短パスとは，図 4 で示すような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．

表 3　副作用関係ペアの素性

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 1，逆であれば 0.
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.
ペア間係り受け最短パス	医薬品と副作用症状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医療表現.

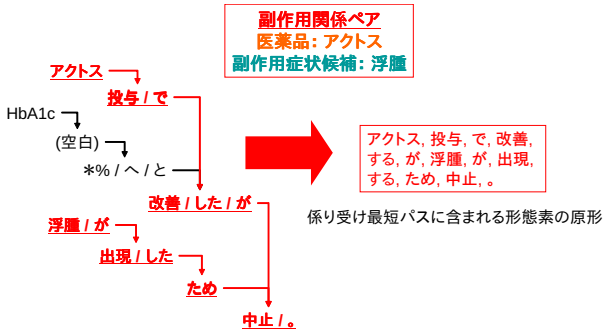

図 4　ペア間係り受け最短パス素性の例

表 4　副作用関係ペア数

	正例	負例	計
評価実験に用いたペア (ペア間係り受け最短パス素性有り)	381	7,487	7,868
同一文中の全ペア	407	9,835	10,242

ステップ 3: 副作用関係の学習

正負のラベルと副作用関係ペアの素性を用いて，副作用関係を Support Vector Machine (SVM)[6] で学習する．

4. 評　　価

提案手法の性能を確認するため，副作用関係コーパスを用いて評価実験を行った．

実験に用いた副作用関係ペア

副作用関係コーパスから，同一文中に含まれる 10,242 副作用関係ペアを抽出した．10,242 ペアのうち，2,374 ペアを含む文でペア間係り受け最短パス素性が得られなかったので，これらのペアは評価実験から除外した．表 4 にペア数と正例・負例数を示す．

実験に用いた素性

ペアの素性には表 5 の組み合わせを用いた．形態素距離，ペア間形態素の抽出には，形態素解析器 JUMAN version 6.0☆を用いた．ペア間係り受け最短パスの抽出には，係り受け解析器 KNP version 3.0☆☆を用いた．

☆ http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
☆☆ http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

表 5　実験に用いた素性

セット	素性
A	文字距離，形態素距離，出現順序，ペア間形態素
B	A ＋ペア間係り受け最短パス
C	B ＋文中の医療表現，副作用症状候補中の医療表現

表 6　最も高い F 値とそのときのパラメータ値

素性セット	パラメータ	適合率	再現率	F 値
A	$\log(\gamma) = -5$, $\log(C) = 2$, $p = 0.10$	29.86	37.54	32.67
B	$\log(\gamma) = -5$, $\log(C) = 2$, $p = 0.10$	39.11	36.57	37.71
C	$\log(\gamma) = -7$, $\log(C) = 3$, $p = 0.15$	37.52	44.42	40.39

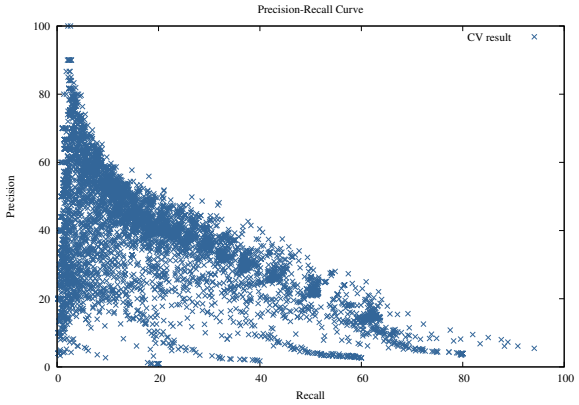

図 5　Precision-Recall カーブ

評価実験 1

各種パラメータを，一定の範囲内で組み合わせた際の精度 (適合率，再現率，F 値) を測定した．SVM の実装には libsvm version 2.89☆を用い，カーネル関数には Radial Basis Function (RBF) を用いた．また，SVM 出力の確度を得るため，libsvm の確率値推定 (probability estimates) オプションを有効にした．

RBF カーネルの γ パラメータの範囲として $[2^{-20}, 2^{0}]$ を用い，SVM の C パラメータの範囲として $[2^{-10}, 2^{10}]$ を用いて，441 通りの γ と C の組み合わせで SVM を学習させた．評価時には，確率値の閾値 p を $[0.05, 0.95]$ の範囲で設定して，5 交差検定で精度を測定した．結果として，各素性セットで表 6 のパラメータ値のときに最も高い F 値が得られた．また，参考までに素性セット C での Precision-Recall カーブを図 5 に示す．

評価実験 2

各種パラメータを，一定の範囲内でのパラメータ探索により自動決定した際の精度を測定した．これは 2 重に交差検定を行うことにより実現した．以下，その手続き

☆ http://www.csie.ntu.edu.tw/~cjlin/libsvm

表 7　パラメータ自動決定時の精度

素性セット	適合率	再現率	F 値
C	32.83	46.22	38.21

表 8　狭義の副作用と広義の副作用の抽出精度

種類	適合率	再現率	F 値
狭義	39.81 (123/309)	56.42 (123/218)	46.68
広義	28.47 (39/137)	23.93 (39/163)	26.00

を示す．

(1)　データから，n 通りの $n-1:1$ の比率の学習・テストセットを作成する．

(2)　n 通りの学習セットそれぞれで，パラメータ γ，C，p の全ての組み合わせで交差検定を行い，平均して最も高い F 値が得られる，n 通りのパラメータの組み合わせを特定する．

(3)　n 通りの学習・テストセットそれぞれを，(2) で得られた対応する γ，C，p で学習・評価し，n 通りの精度の平均を求める．

$n = 5$，素性セット C で，評価実験 1 と同じ SVM の設定を用いて精度を測定したところ，表 7 の結果が得られた．

5. 考　　察

提案手法の現状を確認するため，自動抽出された副作用関係のエラー解析を行った．エラー解析の対象としては，最も高い F 値が得られた表 6 の素性セット C の結果を用いた．

狭義の副作用と広義の副作用

2.1 節で述べたように，副作用関係コーパスでは副作用症状候補を狭義の副作用と広義の副作用に分けている．これらの抽出精度は，エラー解析対象データでは表 8 の値が得られており，狭義の副作用の精度が広義の副作用の精度を大幅に上回った．これは，狭義の副作用の方が数が多いため，また広義の副作用は組み合わせで副作用を表すことがあるためだと考えている．

副作用症状候補の組み合わせによる副作用表現

図 6 に SVM が学習できなかった例を示す．この例では，“溢水” と “増強傾向” の組み合わせで副作用を表現している．“溢水の増強傾向” を 2 つのタグで表現しているアノテーション仕様の妥当性の問題もあるが，現在の SVM の素性には語や句の修飾関係を考慮するものはなく，学習は困難であったと考えられる．

補足的な症状表現の誤判定

図 7 に SVM が誤って副作用関係にあると判定した例を示す．この例の “後腹膜血腫” は “フェロ・グラデュメット” の副作用ではなく，内服に関しての注意事項である．SVM が誤ったのは，医薬品の内服開始の直後に現れる症状は副作用であることが多いためだと考えている．実際には，“後腹膜血腫” は “フェロ・グラデュメット” の内

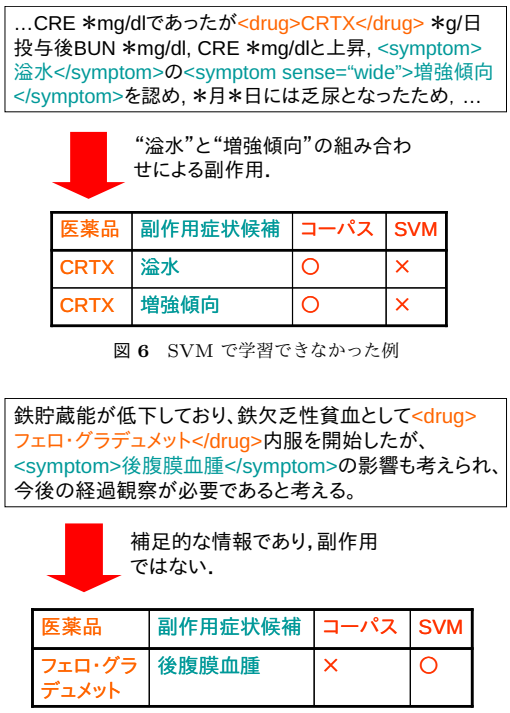

図 6　SVM で学習できなかった例

図 7　SVM が誤った例

服により生じているのではなく，それ以前から存在している．しかし，SVM の素性には発生日時・順序を考慮する情報は含まれておらず，学習は困難であったと考えられる．

6. ま と め

本研究では，医療テキストからの副作用関係の自動抽出手法を提案した．評価実験により，2,577 件の退院時サマリ中で同一文中に現れる副作用関係の抽出精度を測定したところ，最適なパラメータ値で F 値 40.39 が得られ，パラメータ自動決定で F 値 38.21 が得られた．また，副作用関係を狭義と広義の 2 種類に分けた場合，狭義の副作用の抽出精度が広義の副作用の抽出精度を大幅に (F 値で約 20) 上回っていることが確認できた．今後は抽出性能の向上と共に，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に取り組む予定である．

7. 今後の課題

医療知識源の活用

提案手法では，医療辞書，医療オントロジー等の医療知識源の情報は一切活用していない．医療知識源をうまく活用できれば，形態素解析の精度向上や医療用語の抽象化により，副作用関係の抽出性能を向上させられると考えれている．例えば，“倦怠感” と “脱力感” という 2 つの疾患名は，医薬品規制用語集の MedDRA/J☆では同一の上位階層 “無力症” に属する．このため，医療テキストに出現する語を MedDRA/J へマッピングできれば，語間の意味的な関係を素性として利用できるようになると考えている．

より深い構文情報の活用

評価実験では係り受け解析結果を利用した素性を追加することにより，副作用関係の抽出性能が向上した．しかし，現在の素性には最短パス中に出現する形態素という，非常に限定された構文情報しか含まれていない．5 章で述べた，現在の素性では対応が難しい副作用関係を抽出するために，今後は動詞項構造や部分木構造等のより深い構文情報の利用を考えている．

不均衡データ (Imbalanced Data) への対応

評価実験に用いた副作用関係ペアは，正例より負例が大幅に多い不均衡データであった．評価実験では，SVM の確率値推定と確率閾値パラメータ p を導入することにより簡易的に不均衡データに対応したが，根本的な対応策ではないと考えている．今後は，負例の削減手法，不均衡データに適した機械学習手法の導入等を検討している．

同一文中に現れない副作用関係への対応

評価実験では同一文中に現れる副作用関係のみを対象とした．しかし，実際の医療テキストには複数文に渡って記述される副作用関係もある．実際に，評価実験で用いた 2,577 件の退院時サマリ中には 768 ペアの副作用関係が存在しており，同一文中のペアは全体の 53.0%(407/768) である．複数文にまたがる副作用関係の抽出は，文間の関係等を考慮する必要があるため困難であると考えているが，今後はこれらの抽出にも取り組む予定である．

☆ http://www.sjp.jp/~jmo_new2006/php/indexj.php

参 考 文 献

1) Grishman, R. and Sundheim, B.: Message Understanding Conference - 6: A brief history, *In Proc. of COLING 1996*, pp. 466–471 (1996).
2) Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S. and Weischedel, R.: The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation, *In Proc. of LREC 2004*, pp. 837–840 (2004).
3) Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P. and Yuret, D.: SemEval-2007 task 04: Classification of Semantic Relations between Nominals, *In Proc. of SemEval-2007*, pp. 13–18 (2007).
4) Krallinger, M., Leitner, F., Rodriguez-Penagos, C. and Valencia, A.: Overview of the protein-protein interaction annotation extraction task of BioCreative II, *Genome Biology*, Vol. 9(Suppl 2):S4 (2008).
5) 荒牧英治, 三浦康秀, 外池晶嗣, 大熊智子, 増市博, 大江和彦: 退院サマリ可視化システムの構築, 言語処理学会第 15 回年次大会, pp. 348–351 (2009).
6) Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer (1995).