原著論文

闘病ブログに出現する（薬剤名，対象，効果） で表される薬剤服用情報の抽出†

北嶋 志保*・ジェプカ ラファウ*・荒木 健治*

患者が病気や治療方法を正しく理解し、判断、決定するために、医療情報が必要である。そこで我々は、患者の医療情報入手の助けとなるシステムの開発を目指している。本論文では、ナラティブ情報である闘病ブログのスニペット、全文に対し（薬剤名・対象・効果）で表される薬剤服用情報を、手がかり語と構文情報を用いて抽出する手法の提案を行った。既存の抽出パターンと組み合わせることにより、最大24.8ポイント再現率が向上した。薬剤に関する既存の文書から対象単語辞書を作成し、フィルタとして用いることで適合率が最大21.1ポイント向上した。

キーワード：テキストマイニング，医療情報，情報抽出，ブログ，薬剤効果

1. はじめに

治療を進めるにあたり、近年重要視されているインフォームド・コンセントとは、「患者が十分な説明を受け、理解し、納得した上で行う同意」を意味する。しかし、病気や診断方法、治療方法は複雑であるため、突然病気を患った患者がこれらを正しく理解し、判断、決定することは容易ではない。そのため、患者の理解、判断を助け、その際生じる様々な疑問を解決するために医療情報が必要とされている。患者が必要な情報を取得し、それらに満足できた場合、意思決定や治療に対する積極性が促され、QOL(生活の質)や健康状態により良い影響をもたらすことができる。

患者の医療情報入手先に対する調査[1]では、入手先として、医療関係者が最も多く信頼できる情報源であると評価されている。しかし、よりよい意思決定を行うためには、集団に対する効果等が確率的に表される「エビデンス情報」と、体験者の個別的な語りである「ナラティブ情報」の両方が必要である。患者は、同病者から体験談を得ることで、不安が軽減したり、励まされることがある[2, 3]が、医療関係者がこのナラティブ情報を十分に提供するのは困難である。また、担当医師や看護師に立場の違いから質問しにくいといった調査結果[4, 5]からも、患者が情報不足感を抱いていることがわかる。

本研究では患者が医療情報、特にナラティブ情報である同病者の情報入手を助けるシステムの開発を目指している。本手法ではリアルタイムに大規模データが取得可能なネット上に存在する情報に着目し、患者やその家族である闘病者によって書かれたブログ記事を情報源とすることで実体験に基づいた情報を収集する。

日本では個人が匿名で情報を発信することが多い[6]。匿名の情報の信頼性は確かではないが、匿名で発信することにより体裁を気にすることなく、身近な人には知られたいくないような情報も公開することができる[7]。そのため、ネット上から得られる大規模かつ即時的な情報は、既存の調査を上回る可能性がある。また、患者自身の疾患や医療に対する認識、信頼性は不完全だが、多数の情報を収集することにより、同様な情報の有無などから統計的に匿名情報の信頼性を把握することが可能であると考えられる。以上のことから、書き手の率直な意見や感想、悩みと言った情報がウェブから入手可能であり、膨大な情報を利用することにより匿名の信頼性の低さを補うことができると考えている。

このようなシステムにより、患者の性別、年齢などの分布傾向や治療法についての情報だけでなく、未知の薬剤効果など、医師が認識していないような新事実の把握も可能であると考えられる。さらに、体験者から得られる医療情報は、患者にとって治療法などを決定する際の判断材料となるだけでなく、精神的な支えにもなり得る。

ナラティブな医療情報入手システムを構築するため、まず本論文では薬剤の服用による効果や変化に関する記述を、ブログから(薬剤, 対象, 効果)の三つ組で抽出する。薬剤効果は「効く」「効かない」「眠れる」「辛い」

† Blog Based Drug Effects Extraction System Using Drug Name, Object of Change, and Effect Triplets
Shiho KITAJIMA, Rafal RZEPKA and Kenji ARAKI

* 北海道大学情報科学研究科
Graduate School of Information Science and Technology,
Hokkaido University

といった一単語で極性をもつ単語で表されることも多い。しかし、「なくなる」といった評価表現では、「髪がなくなる」と「痛みがなくなる」のように対象によって極性が大きく異なる。このような単語の出現頻度だけでは対応しきれない問題を解決するため、対象と効果で対となる薬剤効果の記述を抽出する。

2. 関連研究

2.1 医療情報検索システム

2.1.1 ナラティブ情報入手サービス

同病者の情報交換手段として、友の会への加入などが一般的に考えられるが、インターネットを利用した情報提供も行われている。乳がんや認知症などのナラティブ情報を交換するサイトとしてDIPEX[8]がある。約150人の患者がインターネット上に顔を出して体験談を語っており、信頼性の高い情報を得ることができる。しかし、日本人は匿名で情報を発信することが多いため[6]、さらに多くの体験談を得ることは難しい。

本論文が対象としているブログは、匿名で気軽に情報発信が行えるため信頼性は低い。しかし、1に述べたように、匿名だからこそ発信できる情報や、直接当事者から手に入れることは難しいような情報を多数集めることができる[7]。

2.1.2 医薬品検索サービス

医薬品の効能や副作用などを検索できる医薬品検索サービスは、企業や法人などにより多数提供されており、医薬品添付文書検索サービス[9]がその一つである。医薬品添付文書とは、各医薬品の製造、販売元が作成した、医薬品の効能や成分、副作用等が記載されている文書である。医薬品添付文書検索サービスでは、薬剤名から添付文書の検索、警告情報、禁忌情報、副作用情報の検索が可能である。このような医薬品検索に関するサービスは、既知の情報を用いているため、関係している可能性のある未知の薬剤効果については検索することができない。

本論文では、患者によって書かれたブログを対象とすることにより、未知の副作用や効能を発見するために必要な情報も取得可能であると考ええる。

2.2 薬剤効果記述箇所特定システム

医薬品検索システムを構築する際には、報告や添付文書から薬剤効果に該当する箇所の特定が必要となる。しかし、自由に記述された文章から、そのような特定作業を手で全て行うには多大な労力が必要である。そこで、近年、自然言語処理技術を中心とした情報技

術の医学分野への応用研究が進められている。

大熊ら[10]は退院時サマリーから副作用について記述されている箇所を特定するシステムの作成を行った。医薬品と副作用症状候補間の距離や構文情報、医療表現などを素性としてサポートベクターマシンの(SVM)[11]を用いた特定を行う。大熊らの研究のように、副作用関係抽出はSVMによりされることが多いが、SVMによる精度は学習データに依存するため、自然言語で書かれた文書に対する十分な抽出精度が得られていない。篠原ら[12]はSVMを用いず構文パターンを用いた薬剤副作用情報の抽出手法を提案した。

電子カルテはICD10やMeSHの用語を用いて記述されることが多く、これらの専門用語辞書を利用した医療情報の特定が進められている。しかし、本論文で対象としているブログ記事の筆者は専門家ではなく使われる用語も一般的なものであるため、専門用語辞典を用いた抽出では不十分である。

そこで本論文では、専門用語辞書を用いない薬剤情報の抽出、また、一般人へ向けで作成された薬に関するHPを利用したフィルタリング方法の提案を行う。さらに、機械学習を用いない抽出方法を考察することにより、医療情報が書かれたブログ記事の特徴を把握することも目的としている。

2.3 商品や場所についての評価を抽出する研究

利用者から得られる医薬品情報を、薬に対する評価情報・意見情報と捉えることもできる。

立石ら[13, 14]の研究では、Webページを対象としており、あらかじめ用意した評価表現辞書を参照して、近接演算により対象物と評価表現を含む一定範囲の文字列を、対象に対する意見として抽出する。その後、パターンマッチングを用いて構文的に意見を表しているかどうかを判定する。抽出に用いる評価表現辞書は、人手で作成したものであり、ドメインに大きく依存するため、ドメインが変わるたびに辞書を作り直さなければならず、辞書作成のコストが高いという問題がある。さらに、立石らの研究は、評価表現にラベル付した「肯定・否定」によって、抽出した意見の極性判定を行う。そのため、対象や属性によって極性が左右される可能性のある、「大きい」「小さい」といった単語が評価表現となる場合を考慮していない。松澤ら[15]の研究では、利用者が「肯定」・「否定」・「無属性」のラベルを付与した評価表現辞書を作成し、その評価表現と利用者が入力したキーワードが、あらかじめ用意したパターンマッチングルールを満たしている場合、キーワードの前後200文字を意見として抽出する。抽出に要する辞書は利用者が作成するため、利用

者の嗜好に合わせた情報抽出が可能となるが、具体的な目的なしに、漠然とあるキーワードについての意見抽出を行いたい場合、辞書の作成は利用者にとって大きな負担となる。

意見の主要部分を(対象, 属性, 評価)の三つ組で表現できるとし[16], ある対象についての意見や評判を三つ組で抽出する研究も多くされている。

酒井ら[17]の研究では, 口コミ掲示板を対象とし, 辞書を用いた評価情報の抽出手法を提案している。評価表現辞書に含まれる単語を含む一文を意見文として, 「(対象)の(属性)が(評価表現)」のパターンを満たし, かつ「(対象)の」が「(属性)が」に係り, 「(属性)が」が「(評価表現)」に係るもののみを抽出している。土田らはブログから対象・属性・評価で表される評判情報を, 確信度を付与してランキングする手法の提案を行った[18]。評価表現辞書に含まれる表現とセットになる属性単語をSVMにより抽出し, 単語間距離により対象と属性を対応付ける。杉木ら[19]は, 辞書を用いず, 自然言語で記述された検索クエリに対応する商品を, 商品レビューテキストとクエリを意見対象・項目・値の三組へと変換することで検索するシステムを作成した。

商品や場所についての評価を抽出する研究では, 評価の対象が明らかであり, 使われる評価表現も限られているため, 評価表現辞書を用いた抽出が可能である。また, 目的や対象が明確である商品レビューなどとは異なり, 本研究が対象とするブログは, 薬剤評価のためだけに書かれたものではないため, 日々の記録から医薬品に関する評価部分をいかに抽出するかが課題となる。また, 薬剤について記述する際使われる評価表現は, 表品レビューなどで用いられるものだけではなく, 「パンパン」や「キリキリ」といった擬態語も多い。擬態語など口語で多く用いられる表現を網羅した評価表現辞書は存在せず, 医療用の評価表現辞書を作成するには多大な労力が必要となる。

2.4 既存研究の問題点と提案システムについて

本論文が提案するシステムは, 関連研究の3つの問題点を解決することができる。第一に, 専門用語を用いず, くだけた表現で記述されたテキストを処理することができるため, 効果や様態を記述する際に多く用いられる擬態語, 擬音語の抽出も可能である。第二に, 大熊ら[10]や篠原ら[12]は薬剤の負の影響である副作用に焦点を当てた抽出手法を提案したが, 本研究の提案システムでは正と負どちらの薬剤効果についても抽出することができ, 抽出された情報は, 未知の薬剤効果を含む可能性がある。第三に, 商品レビューの

ような目的や対象が明らかであるテキストではない, 複数の話題を含むブログ記事からも薬剤効果の抽出が可能である。

3. 対象とするブログとその傾向

3.1 対象とするブログ

ブログは書き手によって様々な記述方法が取られ, 形式も決まっていないため, 必要とする情報が書かれた部分を自動判定することは難しい。そのため, 本研究では, 株式会社イニシアティブが運営する, 闘病者のための闘病体験情報を共有するサービスであるTOBYO[20]に登録された闘病ブログを対象とする。

登録されたブログのうち, ランダムに選出した100件について調査したところ, 利用されているブログサイトの内訳は以下の通りであり, 登録されたブログの形式は多様であることが確認できる。

・アメーバブログ ¹	76サイト
・FC2ブログ ²	7サイト
・ライブドアブログ ³	7サイト
・seesaaブログ ⁴	3サイト
・Yahoo! ブログ ⁵	3サイト
・エキサイトブログ ⁶	2サイト
・ココログ ⁷	2サイト
・その他(自作HP)	1サイト

また, これらのブログサイトのプロフィール情報を調査したところ, 本名の可能性があるファミリーネーム・ファーストネームから成る名前が筆者名として用いられているサイトは1つのみであり, 残り99サイトの筆者の名前はすべてニックネームであった。プロフィールに記述された情報の真偽は定かではないが, 性別は72サイト(男性8, 女性64), 地域は35サイト, 年齢は34サイト, 職業は30サイトで公開されていた。この結果からも, 日本語で記述されたブログ記事は匿名性が高いことがわかる。闘病ブログをTOBYOに登録する際, 著者の病名, 性別, 発病時の年齢などが入力されている。

TOBYOに登録されていない闘病ブログも多数存在するが, それらのブログには性別などのタグ付がされておらず, ブログがどのような人物によって書かれたかなどの情報を抽出する技術も必要となるため, 本論文ではあらかじめ情報が付与されているTOBYOに登

1 <http://ameblo.jp/>

2 <http://blog.fc2.com/>

3 <http://blog.livedoor.com/>

4 <http://blog.seesaa.jp/>

5 <http://blogs.yahoo.co.jp/>

6 <http://www.exblog.jp/>

7 <http://www.cocolog-nifty.com/>

録された闘病ブログを対象とした。また、TOBYO辞典を用いることにより、闘病者によって書かれたブログが収集可能であり、このようなノイズの少ないデータは本システムのアルゴリズム考案に有効であると考ええる。

また、TOBYOではTOBYO辞典を用いて、病名・薬品名・治療方法・医療機関名など固有名詞をキーワードに闘病ブログを検索できる。本論文では、296種の薬剤名についてTOBYO辞典を用いて検索し、検索に用いられた薬剤名(対象薬剤名)の服用情報を、検索結果として得られるブログの要約文であるスニペットから抽出する。

3.2 TOBYOデータベース

本論文では、TOBYO[20]に2011年11月時点で登録されていた闘病ブログ 32,679サイトを対象とする。TOBYO辞典で296種の薬剤名について検索をかけ、得られた闘病ブログスニペットと、その検索元の対象薬剤名、ブログにあらかじめタグ付けされた病名、性別、発病時の年齢、ブログタイトル、ブログ記事タイトル、URLをセットにしてデータベースを作成した。データベースには、検索により得られた103,917データのうち、スニペット内容が重複したデータを除くと全80,745データが格納されている。

闘病ブログは随時更新されているが、TOBYO辞典で検索結果を表示するまでにはタイムラグがあり、すべての検索結果をその都度リアルタイムに入手することは、多大な時間が必要となる。そのため、抽出手法の提案を目的とした本手法では、作成したデータベースを利用する。

3.3 TOBYOスニペットの傾向

3.3.1 薬剤名とその評価の出現位置

TOBYOデータベースからランダムに収集した1,000スニペットを対象とし、対象薬剤名とその服用による

変化や効果、対象薬剤に対する意見などの記述が闘病ブログではどのように出現する傾向があるかの調査を行った。それらの内容には、三つ組の服用情報として表すことができないものも含む。対象薬剤名とその服用による意見を発見できたものが254スニペットあり、そのうち同一文中に薬剤名も存在したものは199スニペットであった。また、対象薬剤名を含む文の前後一文中に意見が存在したものは35スニペット、同一文中でも前後一文中でもないがスニペット中には意見が存在しているものは20スニペットであった。これら意見の一部を表1に示す。

このことから、TOBYOスニペット中では、薬剤服用による意見を含むスニペットの約78.3%で、同一文中に対象薬剤名を記載していることが確認された。そのため、本研究ではスニペット中薬剤名を含む文を対象とした薬剤服用情報の抽出を行う。薬剤名を含まない文に対する関連性の有無の判定や、情報抽出については今後の課題とする。

3.3.2 手がかり語

新たにTOBYOデータベースからランダムに収集した2,000スニペットを対象とした。薬剤名と、(薬剤、対象、効果)の三つ組で表すことのできる服用情報が同一文中に存在する文を第一著者が調査し、それぞれの三つ組要素をタグ付した。その結果、服用情報を含む参考データとして181文を入手した。

抽出した181文の一部を表2に示す。表2(1)~(3)にあるように、服用情報を記載した一文中で、対象薬剤名の後に特定の服用・使用表現(表2の下線部)を使用した後、対象部位や性質の変化を記述している文が138文(76.2%)存在した。138文中に存在した服用・使用表現を手で収集し、取得した32語の名詞・動詞を手がかり語とする。なお、動詞は活用により表層形が異なるため、原形を登録している。抽出した32語の手がかり語として用いる服用・使用表現の一部を表3に示す。

表1 薬剤名とその効果の出現位置例

関係性	薬剤名	スニペット
同一文中	プレドニン	残暑お見舞い申し上げます。暑すぎです…。MSの方々ウートフに苦しんでおられるかと思えます。心中お察しいが変わりないかどうかわいてもらう以外に、もう一つあった。プレドニンを飲み始めてから、どうもコレステロールの数値が高いのでしていた栄養指導を、仕方なく受けるためだ。先生は、どうもプレドニンではなく、食事の影響だと思っているのだ。ところで、今回
前後一文	漢方	漢方を飲んで2ヶ月目。すぐに効果があったのは・・・オリモノが多い→減りましたねへ。これが一番最初が多いのか少ないのかってわからなかったんですね。でも漢方を飲んですぐに量が減ったんです。以前はオリモノシートが年中いつでも乾くのは体質だとばかり思っていたんですが、漢方を飲み始めて1ヶ月たち、途中経過で薬剤師さんに『唇の乾き
その他	リュープリン	最近、寝ている間にやたらと体がほてって目が覚めて熟睡できない。肩凝りもひどくなったし、心臓が突然バ読んで理由がわかった。手術前に生理を止めるために打ったリュープリンの副作用みたい。ほとんどの症状に心当たりがあったので、

表2 薬剤情報としての三つ組を含む文の例

(1)	それからアレドニンを飲み始めたのですが、アレドニンの副作用で骨が弱くなり
(2)	私もおタキ様（タキソール）の後は、肩や背中がすごく凝ります
(3)	現在の発作は、アレピアチンの注入時間前後(6時、14時、22時)に痙攣が出現している
(4)	多分ジブレキサで頭がボケ気味なんだと思います

表3 手がかり語例

副作用	飲む	使用
おかげ	始める	服用
せい	増える	注入
影響	打つ	投与
戻す	変える	変更

表4 対象要素と効果要素間に存在する助詞

助詞	出現回数	割合 (%)
が	104	57.5
は	23	12.7
を	20	11.0
に	7	3.9
にも	4	2.2
も	2	1.1
(その他)	3	1.7
(なし)	18	9.9

3.3.3 対象と効果をつなぐ助詞

2.3.2により得られた服用情報を含む181文を対象とし、三つ組要素である(対象)と(効果)を含む文節が、どのような助詞を用いて関係付けられているかを第一著者が調査を行った。その結果を表4に示す。例えば、表2(1)の文では、「骨が」が「弱くなり」に助詞「が」によって関連付けられている。同様に表2(2)～(4)においても助詞「が」が(対象)と(効果)を含む文節の関連付けに用いられている。

4. システム概要

本論文では患者やその家族によって書かれたブログ記事のスニペットから、(薬剤名, 対象, 効果)の三つ組で表される薬剤の服用情報を、手がかり語と構文情報を用いて抽出する手法を提案する。ここで、対象とは薬剤名の服用もしくは使用により変化・効果が引き起こされた、あるいは引き起こされ得る部位や性質、効果とは評価表現を含む効果・変化のさまを示す表現

図1 システム概要と抽出例

や態度を示している。

システムの構成要素である解析文生成部、三つ組要素存在文節発見部、三つ組要素生成部について以下で説明する。本研究の既存研究との相違点は、我々のシステムは構文情報と特定の手がかり語を用いた三つ組抽出を行うということである。システムの概要と抽出例を図1に示す。

4.1 解析文の生成

TOBYO辞典を用いて取得されたスニペットを、4ステップで変換する。まず、句点を区切り文字として文単位に区切り、1文中に対象薬剤名を含む文を対象とした服用情報抽出を行う。単語情報の参照先であるIPA辞書[21]に存在しない薬剤名が誤って分割され、構文解析ミスにつながることを防ぐため、解析を行う際には薬剤名を1単語の固有名詞「MEDICINE」に置き換える。この置き換え処理は、296種の薬剤名リストに基づき自動で行っている。なお、この処理に形態素解析結果は用いていない。

また、構文解析ミスを防ぐために、対象薬剤名を含まない括弧を削除する。この結果、表2(3)の「(6時、14時、22時)」は削除され、(2)の「(タキソール)」は残される。対象薬剤名が含まれていなくても括弧中に重要な情報が存在する可能性があるため、削除しても差し支えない括弧かどうかの判定をしたのち削除することは今後の課題である。

図2 提案システムの抽出パターン

4.2 三つ組要素存在文節の発見

経験的に仮に作成した図2に示す手がかり語を用いた係り受けパターンにより, 三つ組の要素が存在する文節の位置の特定を行う。図中の長方形は文節を, 矢印は係り受け関係を示しており, いずれもCaboCha[22]による解析の結果得られたものである。抽出ノイズを減らすため, ブログにおける服用表現出現時の特徴から, 各文節末尾に付属する助詞の条件付けを行う。対象要素存在文節と効果要素存在文節をつなぐ際, 助詞「が」「は」「を」が用いられる割合が合わせて81.2%と高かったため, それら三つの助詞に限定した。また, 格助詞(例: で, から, より)は動作・状態の継続や経過に関する意味を持ち, 接続助詞(例: が, ので, て, けれども)は原因・動機などの意味を持つため, 手がかり語存在文節の最後尾単語の条件とした。

三つ組要素存在文節を発見する順序は「薬剤名」, 「手がかり語」, 「効果」, 「対象」要素が含まれる文節の順である。まず, 296種類の薬剤名が含まれる文節を探し, 薬剤名のすぐ後ろ, もしくは係り先の文節に3.3.2で定めた手がかり語が存在するかどうかを判定する。存在し, かつ手がかり語が含まれている文節の末尾助詞が格助詞, もしくは接続助詞である場合, 手がかり語存在文節の係り先文節に「効果」要素が含まれていると判断する。その後, 「効果」要素存在文節に最も近くで「が, は, を」を文節の末尾助詞に用いて係っている文節を, 「対象」要素が含まれている文節と判断する。

4.3 三つ組要素の生成

三つ組要素が存在すると判定された文節に対し, 助詞が出現するまでの自立語を連結させて要素とする。

表5 三つ組要素生成例

(1)	リウマトレックスを減らすと肝機能が改善	(リウマトレックス, 肝機能, 改善)
(2)	グリベックを飲み始めてから体毛に白髪が増えたという意見が結構	(グリベック, 白髪, 増える)
(3)	リオレサールの服用によって筋肉の緊張をほぐす	(リオレサール, 筋肉の緊張, ほぐす)
(4)	リンデロンで右肘は曲がらなくなった	(リンデロン, 右肘, 曲がらない)

具体例を表5で示し, それぞれについて説明する。

(1) 助詞が出現するまでの単語を連結させて要素とした一般的な三つ組要素生成例である。

(2) 動詞は原形で抽出する。

(3) 三つ組要素存在文節が連体化の助詞「の」によって修飾されている場合は, その文節を修飾節として付与する。

(4) 「右」と「肘」のように連続する名詞はつなげる。文節中に打ち消しの意味を持つ「ない」が奇数個存在したときには, 必要に応じて要素の最後尾形態素の活用を変化させ, 「ない」を付加し否定形にする。否定表現について前述した181文の調査を行ったところ, 「ない」を用いた否定の評価表現が3, その他の否定表現は用いられていなかった。そのため今回は, 「～ません」や「減量」といった手がかり語による意味の反転は考慮していない。

5. 評価実験

三つ組要素存在文節を適切に発見し, 抽出できるかどうかの評価を行った。なお, 第一著者が作成した正解データによる評価実験結果はすでに文献[23]で発表済みである。

5.1 データ

TOBYO内にある検索ツールであるTOBYO辞典により薬剤名を用いて検索された, 闘病ブログのスニペットのデータベースより, 一文中に薬剤名を含む重複無しランダム2,000文を抽出の対象とする。

5.2 正解データ

客観性を持たせるため, 北海道大学の20代前半の大学生3名(男性2名, 女性1名)・大学院生2名(男性1名, 女性1名)がアノテーションしたデータを元に, 正解データを作成する。新たに取得したスニペットを句点で区切り, 文章中に薬剤名を含む2,000文に対し, (薬剤・対象・効果)で表される服用情報が含まれると判断される箇所について, 各タグを形態素ごとに付与する。

意味的には等しい抽出範囲の幅の揺れを認めれば2

人以上一致していた258の三つ組を正解データとする。

5.2.1 アノテーション結果に関する考察

5人のアノテーションデータについてK値を求めたところ、0.21と低い値であった。

完全一致が少ない原因を分析するため、第一著者が、2,000文のうち200文に対し服用情報のタグ付けを行い、比較した。その結果200文に対し、第一著者が抽出した三つ組数は、評価者5人の平均17に対して32であった。

第一著者と評価者とのアノテーション結果に差が出た大きな原因は、説明不足であったと考えられる。我々が目指すシステムは、現れていない効果も「現れないことが現れた効果」として意味がある情報であり抽出すべきだと考えている。しかし、例えば、「リフレックスのおかげでかゆみがなくなった」という文に対し、（リフレックス、かゆみ、なくなった）とタグ付けした評価者は一人もいなかった。我々のシステムの目的を考えれば、（リフレックス、かゆみ、なくなった）という情報も正解データとしてタグ付けされるべきである。さらに、「私に取って、タキソテルの脱毛率は少々高めです」について（タキソテル、脱毛率、高め）と（タキソテル、脱毛率、少々高め）と抽出する違いがあるなど、抽出範囲の揺れも完全一致率を下げる大きな要因であった。この意味的には等しい抽出範囲の揺れを容認し、一致しているとした場合、評価者5人が抽出した三つ組の異なり総数は200文中53であった。

第一著者とアノテーションが一致しなかった例について挙げる。ブログの筆者が誤った表現を用いていることや、スニペットの文であるため不完全であること（例：「ある程度はまし」を「程度はまし」と記述）が評価者を混乱させていた。また、「今回の数値はプレドニンを飲み始めて風邪をひいた時を除けば最低数値である」について、「風邪をひいた」ことは「プレドニンを飲み始めた結果」であるのか、筆者の意図する意味かどうかを判定することは、人手でも困難でありシステムに判断させるのは難しい。「熱が上がったら病院なのでボルタレンなど適切な処置をしてくれることでしょう」という文に対して、（ボルタレン、熱、上げる）とタグ付けした評価者もいたが、上がった熱を下げる働きをもつのがボルタレンである。この文章から「ボルタレンは熱を下げる薬である」という情報を抽出するためには、表層的な情報だけではなく一般常識を用いた推論が必要となるため、システムに判定させることは難しい。将来的に推論ができるシステムを作りたいと考えているが、推論が必要な情報は本論文で目指

表6 200文アノテーション結果

一致人数	抽出三つ組数	第一著者との比較			
		一致三つ組数	適合率 [%]	再現率 [%]	F値
3人以上	12	12	100.0	37.5	0.545
2人以上	22	20	90.9	62.5	0.741
1人以上	53	30	56.6	93.8	0.706

すシステムの正解データとして不適切である。

アノテーションされた2,000文のうちランダム抽出した200文に対し第一著者がタグ付けし、評価者と第一著者との一致率を計算した結果を表6に示す。主旨の等しい抽出結果を一致していると見なしたとき、2人以上の評価者が一致したタグ付けした結果のF値が最も高かった。したがって、アノテーションされた2,000文のうち、2人以上の一致が得られた258組の三つ組を正解データとした。この正解データは、第一著者の意図するデータとして最も適切だと考えられる。

今後は事前に評価者に対し、システムの目的についてより深く理解してもらい、アノテーションの訓練をするなど、目指すシステムの正解としてふさわしいデータを作成する必要がある。また、アノテーションにより、個人によって重要視する情報に差が見られたため、抽出する服用情報として主要素である三つ組の代わりに、フレーズでの抽出も医療情報提供には有効であると考えている。

5.3 評価に用いる式

評価には式(1)、(2)、(3)で定義される適合率、再現率、F値を用いる。

$$\text{適合率} = \frac{\text{正解抽出三つ組数} \times 100}{\text{抽出三つ組数}} \quad (1)$$

$$\text{再現率} = \frac{\text{正解抽出三つ組数} \times 100}{\text{正解データ数}} \quad (2)$$

$$\text{F値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \quad (3)$$

5.4 比較対象

図3に示す3つの抽出手法をベースラインとした。ベースライン1は既存手法[17]でレビューなどから意見を抽出する際用いられていたパターンである。また、既存手法[18]では意見抽出の際、評価表現に着目した抽出がされているため、ベースライン2、3では一般的な評価表現辞書であるEVALDIC_ver1.0.1[24]

ベースライン1：パターン抽出

例) マーベロンで肌がつつや ⇒ (マーベロン, 肌, つつや)

ベースライン2：評価表現辞書抽出

例) マーベロン飲みはじめて1年経ったけど、肌が綺麗になった
⇒ (マーベロン, 肌, 綺麗) 登録されている評価表現：綺麗

ベースライン3：評価辞書パターン抽出

例) マーベロンで肌が綺麗になった
⇒ (マーベロン, 肌, 綺麗) 登録されている評価表現：綺麗

図3 ベースラインの抽出パターン

図4 抽出規則A

図5 抽出規則B

を用いたパターンマッチングにより抽出する。辞書には5,234単語が登録されている。なお、これらの抽出手法は三つ組要素存在文節の発見部分で比較対象として用い、解析文生成、三つ組要素生成部分では提案手法と同じ処理を行う。

また、提案手法の助詞の条件付けの有効性を確認するために、図4、図5で示す抽出規則A、Bを用いた抽出も行う。抽出規則Aによる抽出では、提案手法の手がかり語に後続する助詞の条件をつけない。抽出規則Bでは、対象要素候補が存在する文節の最後尾の助詞として、予備調査により用いられやすいことが確認された助詞上位4～6位であった「に、も、にも」も条件に加えている。

5.5 結果と考察

抽出結果を表7に、抽出結果例を表8に示す。評価に用いたデータの正解三つ組数は258である。

文献[23]で述べた第一著者が作成した正解データを用いた実験結果より表7の実験結果は数値的には低いが、提案手法の適合率が最も高く、ベースライン2(評価表現辞書抽出)の再現率が最も高いという結果、F値の大小関係は一致している。

これらの実験結果から、今回提案した手がかり語を用いた抽出が、闘病ブログに対して有効であることが確認された。ベースライン2(評価表現辞書抽出)、ベースライン3(評価辞書パターン抽出)の結果から、薬剤名と対象・効果との関係性の限定により、抽出三つ組数が大きく変化することがわかる。さらに、抽出規則A、抽出規則Bの出力結果から、手がかり語に後続する助詞の条件付けが抽出数に大きく起因していることが確認できた。

提案手法の抽出誤りの理由として、大きく2つ考えられる。要素が存在すると判定された場所の誤りによって出力ミスまたは出力が得られないこと(表8の④)、そして抽出された要素が、三つ組の要素として

表7 評価結果

	出力数	正解数	適合率 [%]	再現率 [%]	F-値
手がかり語と構文情報を用いた提案システム	38	16	42.1	6.2	0.108
ベースライン1：パターン抽出	143	2	1.4	0.8	0.010
ベースライン2：評価表現辞書抽出	662	64	9.7	24.8	0.139
ベースライン3：評価辞書パターン抽出	35	1	2.9	0.4	0.007
抽出規則A	53	16	30.2	6.2	0.103
抽出規則B	105	31	29.5	12.0	0.171

表 8 提案システムの抽出例

	解析文	出力	正解
①	考えたくなくても、やることはないし大量のステロイドのせいで頭が冴えちゃって…	(ステロイド, 頭, 冴えちゃう)	(ステロイド, 頭, 冴えちゃう)
②	幸いにも治療が進むにつれて自力で腎臓が動き出れて、そのまま進級ができました校長先生ありがとうでもステロイドの副作用で顔はパンパン	(ステロイド, 顔, パンパン)	(ステロイド, 顔, パンパン)
③	皆様(●'▽')ノ*【ご】*【き】*【げ】*【ん】*【よ】*【う】*遂に今年もあたしにとって憂鬱な時がやって来ました一応、プログラフはそのままですセルセプトをイムランに変更してでの妊娠・出産は、いいという事でしたが・・・るんですけど・・・で、今回その変更しようとしている『イムラン』というお薬・・・その副作用で骨髄機能に悪さをする事	(イムラン, 悪さ, する)	(イムラン, 骨髄機能, 悪さをする)
④	〈喘息でも道は開ける!〉●喘息でも道は必ず開ける!〜漢方で発作がなくなつて確信を持った〜 (7.6/98-S.Uで発作の度に点滴するのはもうこりごりです	(漢方, ×, ×)	(漢方, 発作, なくなった)
⑤	ヒスロンHの副作用で様々なことがあり、これを機にヒスロンHは中止になりましたです	(ヒスロン, こと, ある)	empty

不適切であること(表8の⑤)である。これらについて考察を行うため、2つの分析を行った。

5.5.1 出力誤り分析

提案手法の出力結果のうち、誤りと判定された22文について誤りの種類と原因の調査を行った。

三つ組が抽出されたが誤りであった文のうち3文は、スニペットとして生成される際に不適切に文が結合され、不自然な文であったことが原因であった。表8の③で示すように、句点を用いた文分割が不適切であり、誤りの原因となった文は3文存在した。この誤りを解決するためには、「☆」「～」「……」による分割など、句点以外の要素を考慮した分割が必要である。

また、効果を記述する際、擬態語など多彩な表現が用いられるため、それらが薬剤効果として適切な表現であるかの判定が難しい。しかし、薬剤の変化や効果が現れる対象は、ある程度限定されていると考えられる。提案手法による出力結果を調査したところ、誤りであった22の三つ組中、13組は対象要素が服用情報の要素として不適切であった。表8の⑤がその一例である。「時間、土曜日、こと、これ」などがその対象要素例として挙げられるが、これらは薬剤の変化が起こる部位や性質として不適切、もしくは指示語であり意味的に不十分である。これらのことから、対象要素の適切性を判定することで誤った抽出を防ぎ、適合率を向上させることが可能であると考えられる。

5.5.2 正解データ出力状況分析

手がかり語を用いた提案手法により正しく抽出された16組の薬剤服用情報は、比較手法では抽出不可能なものであったが、提案手法の再現率は6.2%と低い値であった。2,000文の入力データのうち正解データ258

組の記述形式を調査し、提案手法により出力不可能な242組の三つ組について、その原因を以下に述べる。

未登録の手がかり語を用いて薬剤と変化の関係が示されているものが17組、手がかり語が用いられていないものが104組存在した。表8の④は手がかり語が用いられていないために提案手法で抽出不可能であった例である。スニペットの不完全な一文を対象にしたこと、プログが口語表現で記述されているため助詞の抜け落ちが原因で、26組については単純な構文解析では存在する手がかり語を発見することができなかった。助詞の不一致や手がかり語不在という原因の他に、薬剤名存在文節と手がかり語存在文節との間に意味をもたない「という薬を」などの文節がはさまれているために、手がかり語を発見できない場合もあった。手がかり語を増やし、助詞の抜け落ちや解析の妨げとなる表現を考慮した抽出手法を考える必要がある。

また、提案手法のパターンマッチングに当てはまらないため抽出不可能である正解データは242組中134組存在した。134組中142組は手がかり語に後続する助詞が格助詞、接続助詞以外であり、対象要素存在文節の末尾助詞が「が、は、を」以外である組は134組中64組存在した。これらの三つ組には提案したパターンの助詞の条件付けが当てはまらず、抽出が不可能であった。さらに、パターンマッチングが当てはまらない文には「慢性膵炎には漢方薬も有効的」のように「対象→薬剤名→効果」のような出現順で表記された情報や、「高温期続いてくれているのですがデュファストン効果かな～」のように「対象→効果→薬剤名」の順で記述された情報が258組中約46組(17.8%)存在した。

指示語が薬剤名や対象部位の代わりに用いられているため、照応解析などの処理が必要な文もいくつか存在し、より高度な抽出を行うためには意味を考慮した

抽出が必要である。薬剤の併用に関する記述に対し並列関係を考慮した抽出を行うことも、より多くの服用情報を正確に抽出するためには必要である。

手がかり語が存在しない、抽出パターンにマッチしないなどの問題があったために抽出することができなかった文のうちいくつかに対しては、ベースライン1(パターン抽出)、2(評価表現辞書抽出)により正解三つ組が抽出可能であった。そのため、再現率を上げるためには、提案手法と他の手法とを組み合わせたシステムが有効であると考えられる。

6. 組み合わせによる再現率向上手法

より多くの薬剤服用情報を抽出するため、提案手法により三つ組が得られなかった文についてはベースライン2(評価表現辞書抽出)を用いて抽出を行うシステムを作成し、評価を行なった。評価には5で使用したデータと式(1)~(3)を用いる。その結果を表9に、抽出結果例を表10に示す。

正解数が64増加し、再現率は提案手法より24.8ポイント上昇しF値は最も高い値となった。また、組み合わせたことにより抽出できなくなった正解三つ組は存在しなかった。このことから、手法の組み合わせが再現率向上に有効であることが確認された。しかしベースライン2(評価表現辞書抽出)による出力には誤りも多く含まれているため、適合率は下がる結果となった。出力の誤りをなくすために、抽出された要素の服用情報としての適切性を判断する必要がある。

表9 組み合わせシステム評価結果

	出力数	正解数	適合率 [%]	再現率 [%]	F-値
手がかり語と構文情報を用いた提案システム	38	16	42.1	6.2	0.108
ベースライン2: 評価表現辞書抽出	662	64	9.7	24.8	0.139
組み合わせシステム (提案システム + ベースライン2)	681	80	11.7	31.0	0.170

表10 組み合わせシステムの抽出例

解析文	出力	正解
メキシチール飲むようになってから動悸はするけど期外収縮はあんまり感	(メキシチール, 動悸, する)	(メキシチール, 動悸, する)
骨がステロイドで弱くなってきたので、今後のことを考えると、飲んで	(ステロイド, 骨, 弱くなる)	(ステロイド, 骨, 弱くなる)
タキソテールだと白血球さがから私には厳しいです、なによりも	(タキソテール, 私, 厳しいんです)	(タキソテール, 白血球, さがる)
夜の便に、オレンジっぽい箇所が小、こういった副作用はやめて欲しいです	(漢方, 副作用, やめる)	empty

7. 辞書による適合率向上手法

薬剤の服用について記述する際、評価表現は多様だが、変化や効果が現れる対象はある程度限定されていると考えられる。そのため、既存の薬剤に関する効果や副作用が書かれた文書から、対象単語として用いられる可能性がある部位や状態を示す単語を収集し、対象単語辞書を作成した。単語の収集方法は、「接頭詞+名詞」を基本とし、接頭詞はなくてもかまわない。また、連続する名詞についてはつなげたものを一単語として登録する。また、日付や期間といった対象として不適切な単語を除外するため、数字を含む場合は登録しない(例: 1時間)。一般向けに優しい表現で書かれたお薬110番[25]から作成した辞書と、薬剤添付文書から作成した専門用語を多く含む辞書、二つを用いて各手法により抽出された対象要素が適切かどうか判断する。

例えば、抽出された対象要素が「膝の青あざ」であったとする。この対象要素をMeCab[21]により形態素解析し、「膝(名詞)/の(助詞)/青(接頭詞)/あざ(名詞)」を得る。構成要素のうち接頭詞(なくても可)+名詞(連続する場合は結合する)である「膝」と「青あざ」が辞書に含まれているかどうかを検索する。「膝」、「青あざ」どちらかひとつでも辞書に含まれていた場合には「膝の青あざ」が服用情報の対象要素として適当であると判断し、この対象要素を含む三つ組を出力する。

7.1 お薬110番辞書(110辞書)

お薬110番[25]とは薬専門のサーチエンジンである。患者やその家族向けに、専門的な言葉ではなく優しい表現を用いて書き直されているのが特徴である。ウェブページから取得できた1,143ページの薬剤情報のうち、用法についての記載箇所を除く概要、効果、副作用などについての記載箇所から辞書を作成する。その結果、辞書には12,170単語が登録された。その一部を表11に示す。

表11 110辞書登録単語例

てんかん	けいれん 発作	失神発作	不機嫌	躁うつ病
目	副作用	肝機能 検査	のど	意識障害
かゆみ	高アンモニア 血症	集中力	発熱	発疹
食欲不振	気分	嘔吐	意識障害	車
作業	手足	しびれ	毎日	頭痛
躁状態	脳	偏頭痛 発作	物忘れ	眠気
吐き気	青あざ	症状	守り	その他 さまざま

表12 MPI辞書登録単語例

てんかん	焦点発作	混合発作	不機嫌	躁うつ病
視覚異常	副作用	肝機能	腹痛	意識障害
下痢	アンモニア値	血小板減少	発熱	発疹
食欲不振	悪心	嘔吐	言語障害	姿勢
CPK	尿失禁	便秘	月経不順	頭痛
躁状態	脳	偏頭痛発作	健忘	傾眠
嘔気	硬直	症状	厚生省	効果毎

7.2 薬剤添付文書辞書 (MPI辞書)

医薬品医療機器総合機構が運営する医薬品医療機器情報提供ホームページ[9]から、医療用医薬品と一般用医薬品の添付文書情報を用いて辞書を作成する。薬剤添付文書は、各医薬品の製造、販売元が作成して

おり、医学用語が多く用いられているのが特徴である。TOBYO辞典を用いて取得したブログから抽出した三つ組のフィルタに用いることが目的であるため、TOBYO辞典で検索する際クエリとして用いた296種の薬剤名について添付文書情報を取得し、薬効と副作用についての記述箇所から辞書を作成する。その結果、辞書には5,380単語が登録された。その一部を表12に示す。

7.3 辞書の評価実験結果に対する考察

辞書フィルタを用いた抽出結果を表13に、フィルタ例を表14に示す。評価には5で使用したデータと式(1)～(3)を用いる。システムが出力した対象要素を辞書フィルタにかけることにより、提案システムに対するMPI辞書使用時で適合率が最大21.1ポイント増加した。また、表13に示した5種類の手法に対し、フィル

表13 辞書フィルタ評価結果

	出力数	正解数	適合率 [%]	再現率 [%]	F-値
110 辞書フィルタ					
手がかり語と構文情報を用いた提案システム	29	15	51.7	5.8	0.105
組み合わせシステム (提案システム+ベースライン2)	398	65	16.3	25.2	0.198
ベースライン1：パターン抽出	110	2	1.8	0.8	0.011
ベースライン2：評価表現辞書抽出	379	50	13.2	19.4	0.157
ベースライン3：評価辞書パターン抽出	28	1	3.6	0.4	0.007
MPI 辞書フィルタ					
手がかり語と構文情報を用いた提案システム	19	11	57.9	4.3	0.079
組み合わせシステム (提案システム+ベースライン2)	181	38	21.0	14.7	0.173
ベースライン1：パターン抽出	49	0	0.0	0.0	
ベースライン2：評価表現辞書抽出	166	27	16.3	10.5	0.127
ベースライン3：評価辞書パターン抽出	11	0	0.0	0.0	

表14 辞書フィルタ抽出例

入力	出力	フィルタの判定		正解
		110 辞書	MPI 辞書	
ステロイドの使用で関節痛が改善された(少量の維持量で、良い状態を	(ステロイド、関節痛、改善された)	正	正	正
女医先生が、プレドニンを増量してから体調が良くなったように見えると仰った	(プレドニン、体調、良くなったよう)	正	誤	正
たしかにタキソテルの副作用で体全体が浮腫んでいたもので、腕は両方とも腫れて重くて痛いな～とはもっと前から思ってたんだけど、これもタキソテルの副作用だと思ってた	(タキソテル、体全体、浮腫む)	正	誤	正
そういえばアザルフィジン服用直後って副作用で結構しんどかったなあ……これは大丈	(アザルフィジン、これ、大丈)	正	正	誤
2008.8.4骨シンチの結果は…問題なし、遠隔転移はなしでしたやった!!良かった!!これで気持ち的にはそうと次回ECの時、Her2が陽性だと予後が悪いっていうけど、でもハーセプチン使えるし…どっちが良いのかな?微妙?EC2クール目は前回同様	(ハーセプチン、遠隔転移、同様)	誤	正	誤
投与中も胸が変な感じになる原因も分かった・・・パクリタキセル入れる時のアルコール成分がかああっつと胸周りを熱くする	(パクリタキセル、胸周り、熱くする)	誤	誤	正

タ利用前後の適合率を比較したところ, 110辞書で平均5.7ポイント, MPi辞書で平均7.5ポイントの適合率の向上が確認でき, 対象用語辞書を用いたフィルタリングが適合率の向上に効果的であることが確認された. フィルタリングなしの出力では, 対象要素として日時や期間が抽出されている誤り例が存在したが, フィルタリングによりそれらの誤りを除くことができた.

例えば, フィルタを用いた出力では, 対象要素である「朝の眠り」の「眠り」が辞書に登録されているため正しい対象要素として判断する. しかし, 「一ヶ月半」は辞書に無い単語のため, 対象単語として不適切であると判断できる. このように服用情報として不適切な対象要素を除去することで, 適切な三つ組を出力することができる.

2つの辞書を比べると, MPi辞書の登録単語数は少なく, 表現も専門的である. 例えば, 同意義ではあるが110辞書には「吐き気」, MPi辞書には「嘔気」が登録されているといった違いがある. したがって, MPi辞書をフィルタとして利用した際には適切な対象要素も多く除かれてしまうが, 110辞書では適度に不適切な対象要素が除去されるため, 組み合わせシステムに用いた際に最も高いF値となっている. つまり, 話し言葉で書かれたブログにはお薬110番から作成した辞書がより適切であったと考えられる.

また, 辞書へ登録する際, 単語の意味まで考慮していなかったため, 「こと」「の」といったノイズが辞書に含まれていたため, より辞書を有効活用し適合率を向上させるためには, 辞書のデータクリーニングが必要である.

8. ブログ全文に対する抽出

スニペットを抽出対象としたために, 文が不適切につながられ情報として不十分である場合や, 構文解析ミスを引き起こす原因となっている場合が約13.6%(258文中35文)あった. スニペットに起因する問題を回避し, より多くの医療情報を抽出するために, ブログ記事全文を対象とした.

スニペットよりも長く, 医療情報以外の情報も多く含むと考えられるため, 「 $\cdot$ 」「 ω 」「 $((+))$ 」などの文字列はあらかじめ除去する. この操作により顔文字などの情報が消えてしまうが, それらに含まれる感情も考慮した医療情報の抽出は今後の課題とする.

8.1 正解データ作成方法

ブログ記事全文に対する評価実験を行うために, 正解データを作成した. 薬剤名を含む文のうち, 第一著者が薬剤服用情報の可能性がある部分をタグ付けし, その適切性を評価者が判断する. 評価者はすべて北海道大学の20代前半の大学生2名(男性2名)・大学院生3名(男性2名, 女性1名)である.

その結果, 3,000ブログ記事のうち, 545文から抽出された745組の三つ組を薬剤服用情報の正解データとする. この正解データには現れていない効果(例えば「痛みがなくなった」など)も含まれており, 正と負どちらの薬剤効果についても抽出するという本論文の目的に即したデータとなっている. 評価者による評価結果の一致率はK値0.67であり, 実質的に一致していた.

また, スニペットに対するアノテーションの際問題

表 15 全文を対象とした評価結果

	出力数	正解数	適合率 [%]	再現率 [%]	F-値
手がかり語と構文情報を用いた提案システム	52	24	46.2	3.2	0.060
組み合わせシステム (提案システム+ベースライン2)	1412	181	12.8	24.3	0.168
ベースライン1: パターン抽出	193	17	8.8	2.3	0.036
ベースライン2: 評価表現辞書抽出	1386	170	12.3	22.8	0.160
ベースライン3: 評価辞書パターン抽出	58	8	13.8	1.1	0.020
110 辞書フィルタ					
手がかり語と構文情報を用いた提案システム	36	20	55.6	2.7	0.051
組み合わせシステム (提案システム+ベースライン2)	815	141	17.3	18.9	0.181
ベースライン1: パターン抽出	130	14	10.8	1.9	0.032
ベースライン2: 評価表現辞書抽出	795	133	16.7	17.9	0.173
ベースライン3: 評価辞書パターン抽出	40	6	15.0	0.8	0.015
MPi 辞書フィルタ					
手がかり語と構文情報を用いた提案システム	18	10	55.6	1.3	0.026
組み合わせシステム (提案システム+ベースライン2)	420	75	17.9	10.1	0.129
ベースライン1: パターン抽出	60	13	21.7	1.7	0.032
ベースライン2: 評価表現辞書抽出	412	73	17.7	9.8	0.126
ベースライン3: 評価辞書パターン抽出	17	5	29.4	0.7	0.013

であった推論が必要な薬剤情報に対応するため、薬剤服用情報の定義を「薬剤の服用に関わる変化」とし、「ウンチが出ていないのでラキソベロンを飲まなければならなかったけど、少し時間をおいて内服する事になった」には(ラキソベロン, ウンチ, 出ていない)の三つ組をタグ付けした。

8.2 実験結果とその考察

各抽出方法による出力結果を表15に示す。

ブログ全文を対象とした抽出結果は、スニペットを対象とした抽出よりもさらに再現率が2.2ポイント低かった。全文ではスニペットとして表示されない範囲まで対象とされるため、より多くの情報が含まれており、その分多彩な記述表現が存在していたためだと考えられる。「薬の副作用なのか朝が異常にだるいので、薬を追加することでそれが強くなるのが怖く、もう2週間デパスだけで様子を見ることにしました」のような文から「薬」が何を指すのかを、「初回のインターフェロン終了後は数ヶ月のあいだ吐き気などが継続したのですが、今回の治療後はそんなことも一切なく」では「一切ない」のは「吐気」であるなど照応解析処理も必要である。

また、適合率を向上させるためには抽出三つ組の適切性を判断する必要がある。しかし、評価表現辞書抽出により「今骨転移でナベルピン、手足の痺れ、感覚マヒを和らげるため「ピドキサル錠」をDrから処方されてます」から抽出された(ピドキサル錠, 感覚マヒ, 和らげる), (ナベルピン, 感覚マヒ, 和らげる)のように、三つ組要素のみでは判断できない例も多く存在する。そのため、要素としての適切性だけではなく、薬剤名との位置関係や構文情報を用いた判断、さらには意味を考慮した抽出が必要であると考えられる。

9. まとめと今後の課題

闘病ブログのスニペットに対し、(薬剤名・対象・効果)で表される薬剤服用情報を、手がかり語と構文情報を用いて抽出する手法の提案を行った。また、薬剤に関する既存の文書から対象単語辞書を作成し、フィルタリングに用いる事で適合率を向上させた。また、ブログ全文に対しても抽出手法を適用することで、各手法の利点と全文に対する抽出課題を確認した。

今後の課題として、以下の5つが挙げられる。

- ・解析文生成時に、削除して差し支えない括弧かどうかの判定、文へ分割する際の区切り箇所の判定
- ・文末表現である「～ません」や、「減量」といった手がかり語の表現による意味の反転を考慮
- ・手がかり語を増やす

- ・薬の併用に対応する
- ・辞書をより有効利用するために、データのクリーニングを行う

また、アノテーションから、服用情報を要素として抽出するには個人差が大きく、要素の抽出では情報量が不十分な場合もあることが確認された。したがって、服用情報を単語単位ではなく、フレーズ単位で抽出することも効果的であると考えられる。

さらに、抽出した情報に対しそれが伝聞情報なのか体験情報なのか、事実なのか仮定なのかを判定したいと考えている。

TOBYOに登録されていないブログサイトからも今後さらに多くの医療情報を入手するためには、筆者の性別や年齢など個人情報の抽出も課題である。

今後さらに多くの情報を抽出するためには、助詞抜け落ちの補完や、意味解析技術や照応解析技術を用いた指示語の解析など、より高度な言語処理技術が必要であると考えられる。

最終的な目的として、ナラティブ情報を簡単に患者同士でシェアできるようなシステムを作成し、医療情報検索の助けをしたいと考えている。この医療情報とは、すでに把握されている情報だけでなく、新しい情報も含む。患者にとって判断材料や支えとなる情報を提供でき、また、未知の医療情報発見にも役立つシステムを目指している。

オンラインでの健康情報交換が発展しているため、もっと多くの人たちが情報を共有することができれば、医師が発見できないような病気の傾向などを突き止めるチャンスも多いにあると考えている。

参考文献

- [1] 谷口綾, “一期的乳房再建術を受けた乳がん女性の思いと看護支援”, 看護教育 45 (8): pp.660-663, 2004.
- [2] 松本梓, 新垣円, 鄭迎芳, 井ノ口珠喜, 酒井梢, 塩田早恵, 高野学, 高橋都, “医療情報源としてのインターネットに対する乳がん患者の認識—メーリングリスト参加者を対象としたインターネットサーベイから—”, 医療情報学 25 (3): pp.151-159, 2005.
- [3] 前田優雅, 国府浩子, 藤井徹也, “治療中の乳がん患者に及ぼす同病者からの影響と関連する要因—乳がん患者会会員を対象として—”, がん看護 14 (6): pp.711-716, 2009.
- [4] Masahiko Tsuchiya and Sandra Horn, “An Exploration of Unmet Information Needs among Breast Cancer Patients in Japan: A Qualitative Study”, European Journal of Cancer Care, 18 (2): 149-155, 2009.
- [5] 国府浩子, “初期治療を選択する乳がん患者が経験する困難”, 日本がん看護学会誌 22 (2): 14-22, 2008.
- [6] 吉田等明, 村上武, 和崎宏, 五味壮平, “オンラインでの匿名性と倫理観”, コンピュータ&エデュケーション 24, 2008.

- [7] 折田明子, 三木草, 小川美香子, “発信しづらい情報交換における匿名性の効果ーダイエット食品クチコミ調査からー”, 情報社会学会 Vol.2 No.2, pp.114-127, 2007.
- [8] 認定特定非営利活動法人 (NPO) 健康と病いの語りディベックス・ジャパン.
<http://www.dipex-j.org/> (2014/4/7 アクセス)
- [9] 独立行政法人 医薬品医療機器総合機構 医薬品医療機器情報提供ホームページ.
<http://www.info.pmda.go.jp/index.html> (2014/4/7 アクセス).
- [10] 大熊智子, 三浦康秀, 外池昌嗣, 増市博, 篠原 (山田) 恵美子, 荒牧英治, 大江和彦, “医薬品の副作用調査を目的とした統合的言語処理システム”, 言語処理学会第 17 回年次大会 発表論文集, pp.85-88, 2011.
- [11] Corinna Cortes and Vladimir Vapnik, “Support Vector Networks”, Machine Learning, Vol.20, pp.273-297, 1995.
- [12] 篠原恵美子, 服部主悟, 三浦康秀, 外池昌嗣, 大熊智子, 増市博, 荒牧英治, 大江和彦, “構文パターンに基づく薬剤副作用情報の自動抽出と評価”, 医療情報学 31 (Suppl.), pp.521-524, 2011.
- [13] 立石健二, 石黒義英, 福島俊一, “インターネットからの評判情報検索”, 人工知能学会誌 19 巻 3 号, pp.75-82, 2004.
- [14] 立石健二, 石黒義英, 福島俊一, “インターネットからの評判情報検索”, 情報処理学会研究報告, 自然言語処理研究会報告 2001 (69), pp.75-82, 2001.
- [15] 松澤祐太, 鈴木裕利, 石井成郎, 小出周之, “ユーザ嗜好に着目した評判情報の抽出手法に関する提案と評価”, 情報処理学会研究報告, 情報システムと社会環境研究報告 2011-IS-118 (1), pp.1-8, 2011.
- [16] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一, “テキストマイニングによる評価表現の収集”, 情報処理学会研究報告, 2003-NL-154, pp.77-84, 2003.
- [17] 酒井義和, 荒木健治, “反対語を利用した文脈依存評価表現の感情極性判定”, 電子情報通信学会論文誌 D, 情報・システム J93-D (9), pp.1778-1789, 2010.
- [18] 土田正明, 水口弘紀, 久寿居大, “評判検索のための対象, 属性, 評価の 3 項関係のランキング法”, 第 22 回人工知能学会全国大会, 2P2-05, 2008.
- [19] 杉木健二, 松原茂樹, “消費者の意見に基づく商品検索”, 情報処理学会論文誌, Vol.49, No.7, pp.2598-2603, 2008.
- [20] TOBYO::日本の闘病記 44000.
<http://www.toby.jp/> (2014/4/7 アクセス)
- [21] MeCab: Yet Another Part-of-Speech and Morphological Analyzer.
<http://mecab.sourceforge.net/>
- [22] CaboCha: Yet Another Japanese Dependency Structure Analyzer.
<http://chasen.org/taku/software/cabocha>
- [23] Shiho Kitajima, Rafal Rzepka and Kenji Araki, “Performance Improvement of Drug Effects Extraction System from Japanese Blogs”, in the Proceedings of 2013 IEEE Seventh International Conference on Semantic Computing, pp.383-386, 2013.
- [24] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一, “意見抽出のための評価表現の収集”, 自然言語処理, Vol.12, No.3, pp.203-222, 2005.
- [25] ファーマ・フレンド (小林薫), お薬 110 番. http://www.jah.ne.jp/~kako/frame_dwm_search.html (2014/4/7 アクセス).

(2014年 4 月11日 受付)
(2014年10月 2 日 採録)

[問い合わせ先]

〒060-0814 札幌市北区北14条西9丁目
北海道大学 大学院 情報科学研究科 言語メディア学研究室
北嶋 志保
TEL : 011-706-7389
FAX : 011-709-6277
E-mail : shiho@ist.hokudai.ac.jp

著者紹介

きたじま しほ
北嶋 志保 [非会員]

2012年北海道大学工学部情報エレクトロニクス学科卒業。現在、同大学大学院博士後期課程在学中。自然言語処理、特に医療情報に関する研究に従事。医療情報学会学生会員。

あらか きんじ
荒木 健治 [正会員]

1982年北海道大学工学部電子工学科卒業。1988年同大学大学院博士課程修了。工学博士。同年、北海学園大学工学部電子情報工学科助手。1989年同講師。1991年同助教授。1998年同教授。1998年北海道大学大学院工学研究科電子情報工学専攻助教授。2002年同教授。現在、北海道大学大学院情報科学研究科メディアネットワーク専攻教授。自然言語処理、特に、機械翻訳、音声対話処理などの研究に従事。日本知能情報ファジィ学会、電子情報通信学会、情報処理学会、言語処理学会、人工知能学会、日本認知科学会、ACL、IEEE、AAAI 各会員。

ジェプカ ラファウ [非会員]

2004年北海道大学大学院工学研究科電子工学専攻博士後期課程修了。2004年北海道大学大学院情報科学研究科メディアネットワーク専攻助手。2007年同助教。現在に至る。言語学修士、工学博士。常識的知識獲得、感情処理、マシン倫理学、集合知などの研究に従事。電子情報通信学会、情報処理学会、人工知能学会、言語処理学会、日本認知科学会、AAAI、各会員。

Blog Based Drug Effects Extraction System Using Drug Name, Object of Change, and Effect Triplets

by

Shiho KITAJIMA, Rafal RZEPKA and Kenji ARAKI

Abstract :

Medical information is required for patients to understand the risks and benefits of procedures proposed by a doctor, and to select the treatment plan that they prefer. We aim to create a system that helps patients to collect medical information. The purpose of this paper is to extract descriptions of the effects caused by taking drugs as a triplet of expressions from illness survival blogs' snippets and full blog articles including narrative information. This paper proposes a method to extract such triplets using specific clue words and parsing the results in order to extract from blogs written in natural language. The experiments show that recall was improved by 24.8 points when we combined our proposed method and a baseline system, and precision was improved by 21.1 points when we utilized filters using dictionaries we created from existing medical documents.

Keywords : text mining, medical information, Information Extraction, blogs, drug effects

Contact Address : Shiho KITAJIMA

Language Media Laboratory Graduate School of Information Science and Technology Hokkaido University

Kita-ku, Kita 14, Nishi 9, 060-0814, Sapporo, Japan

TEL : 011-706-7389

FAX : 011-709-6277

E-mail : shiho@ist.hokudai.ac.jp