闘病ブログを対象とした 手がかり語を用いた薬剤服用情報の抽出手法

An Extraction Method of Medication Usage Information Using Clue Words from Illness Survival Blogs

○北嶋　志保，荒木　健治，ジェプカ　ラファウ
○Shiho Kitajima, Kenji Araki, Rafal Rzepka
北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

Abstract: It is very useful to constructing system that can predict future situation of illness from medical records written by patients in natural language because it will help to ensure informed consent. As the first step, we propose a method to extract description of the effect caused by taking the medicine as a triplet of expressions - medicine name, object of change, and its effect from illness survival blogs. There is no database of evaluation expressions for medical treatment. Moreover usual extraction patterns are not suitable for processing these blogs. Therefore, we decided to extract a expressions triplet using specific clue words and by parsing results. In the experiments, we have confirmed that medication usage information can be extracted with high accuracy compared to other existing methods.

1. はじめに

近年，ネットワークの普及に伴い，Web上へ患者やその家族が体験した情報を容易に発信することが可能となっている．なかでも，ブログ上で日々の記録として書かれた薬剤や治療法に対する評価や主観的な意見は，同じ病気を持つ患者にとって意思決定の判断材料となり非常に有用なものである．しかし，決められた形式もなく自由に記載された記事から，すべての情報を閲覧し，自分が求める情報を収集することは多大な労力が必要となる．そこで我々は，患者やその家族が書いた闘病体験記である闘病ブログから，医療に関する評価情報を獲得し，それらの評価を数値で客観的に表現し整理することで，求める評価情報の変化を時系列で視覚的にとらえることができるシステムの構築を目指す．このようなシステムが実現できれば，過去に自分と同じ病気を持つ人が，どのような治療を受け，どのような体調の変化が起きたのかを統計的に把握することが可能となる．このような情報は，これから辿る未来の予測を可能とし，患者のインフォームド・コンセントの手助けとなる．

本稿ではこのようなシステム構築の第一段階として，闘病ブログから薬剤の服用による効果や変化についての情報を〈薬剤・対象・評価〉の三つ組みで抽出する手法を提案する．

〈対象・属性・評価値〉で表すことのできる意見や評判を三つ組みで抽出する研究は多くなされている．酒井ら[1]の手法は，評価表現辞書にあ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〉」に係るものの抽出を行う．土田ら[2]の手法は，評価表現辞書にある単語が存在する文とそのひとつ前の文までに含まれる単語を属性候補と考え機械学習により最も〈評価表現〉と対応している名詞を〈属性〉に決定する．また，〈属性〉と〈対象〉の存在距離が50文字以内であったとき，それらを三つ組として抽出する．酒井ら，土田らの辞書を用いた抽出手法は，人手による辞書の構築の負担や，辞書がドメインに依存してしまうという問題がある．それに対し杉木ら[3]は，辞書を用いず係り受けを考慮したパターンマッチングにより属性・評価を抽出する．これらの手法は，どれも商品レビューのような意見が書かれやすいテキストからの抽出を行っており，対象も明らかである．しかし我々が対象とする闘病ブログは，形式ばらない自由な記述が特徴であり，対象も明確ではない．また，医療用の評価表現辞書も存在しない．

2. 提案手法

本手法では，TOBYO[4]に登録された闘病ブログに特徴的に使われる手がかり語と構文情報を用いて，薬剤の服用情報の抽出を行う．TOBYOサイト内にあるブログ検索ツールであるTOBYO辞典を用いて，登録されている296種の薬剤名を検索し，その結果得られた闘病ブログの要約文であるスニペットを対象とする．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

報が書かれた部位を自動判定することが難しい．また，処理時間を考えると全てのブログに対し全文検索を行うことは現実的ではないため，スニペットを対象とした．

TOBYOに登録された闘病ブログは全32,679サイトからTOBYO辞典を用いて296種の薬剤名で検索し得られた各スニペットについて，闘病ブログスニペット，検索元となった薬剤名，ブログにあらかじめタグ付けされた病名，性別などをひとセットとしてデータベースを作成した．TOBYO辞典で結果を表示するまでにタイムラグがあること，TOBYO辞典がテスト版であるため突如使用不可能となる可能性があることから，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．

我々は服用情報の抽出を行なったのち，それらの客観的評価を行うシステムの構築を目指す．「良い」「悪い」「効く」といった，対象の影響を受けないものには客観的評価結果を人手で付与することが可能であり，自動判定の必要性が低い．そのため，対象による極性の変化を考慮できるよう〈薬剤・対象・評価〉で表されるような服用情報の抽出手法を提案する．

TOBYO辞典によりランダムに取得したスニペットに，人手で〈薬剤・対象・評価〉で表される服用情報のタグ付けを行った．その結果，TOBYOスニペットのうち服用表現は78.3%の割合で同一文中に存在することが確認された．そのため，本手法ではスニペット中の薬剤名を含む一文から服用情報を抽出する．処理の流れを図1に示し，各処理について説明する．

２．１．手がかり語

TOBYOスニペットの特徴を解析するため，ランダム抽出した一文中に〈薬剤・対象・評価〉で表される服用情報を人手でタグ付けしたところ，181文に服用情報が含まれていた．その一部を表1に示す．表1(1)～(3)にあるように，服用情報を含む文中で，対象薬剤名の後に特定の服用・使用表現（表の太文字下線）を使ったのち，対象部位や性質の変化を述べている文が138文あり，この全体に占める割合は76.2%であった．これらの服用・使用表現を人手で収集したところ，32語の名詞・動詞を抽出することができた．なお，動詞は活用により表層形が異なるため原形をカウントしている．収集した32語の名詞・動詞を闘病ブログから服用情報を抽出するための手がかり語とする．手がかり語の一部を表2に示し，これらを服用表現と定義する．

２．２．解析文生成

入力スニペットを句点「。」を区切り文字として文単位に区切り，服用情報の対象となる薬剤名を文中に含むものに対し抽出を行う．

係り受け解析のエラーを少なくするために，文中に含まれる不要な括弧とその中身を削除する．不要な括弧とは，括弧中に対象薬剤名を含まないものを指す．また，構文解析に用いるCaboCha[5]は，入力文が体言止めの場合解析を行うことができない．そのため，服用情報抽出には影響のない「です」を文末に付与する．CaboChaの参照するIPA辞書[6]に登録されていないために，誤った形態素に分割されることにより発生する構文解析のエラーを防ぐため，全薬剤名を”MEDICINE”へと置き換える．

２．３．三つ組発見

図2に示す手がかり語と係り受けの出現パターンを考慮した抽出規則により，三つ組要素が存在する

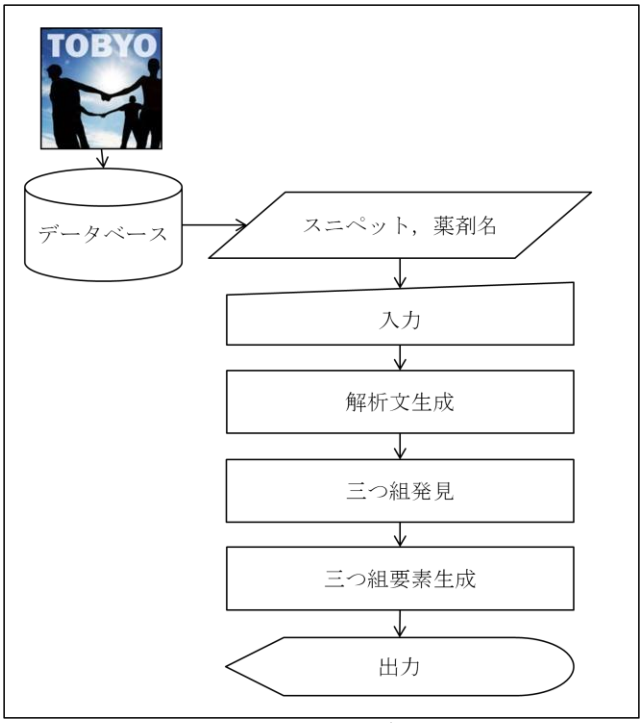


図１　処理の流れ

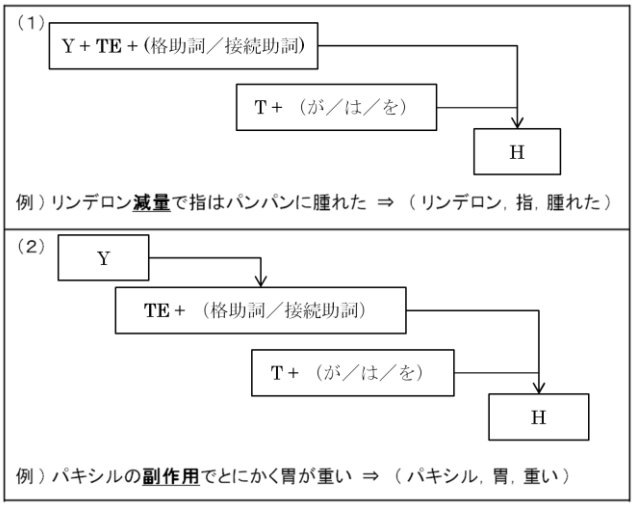
表1　服用情報を含む文の例

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

表2　服用表現の例

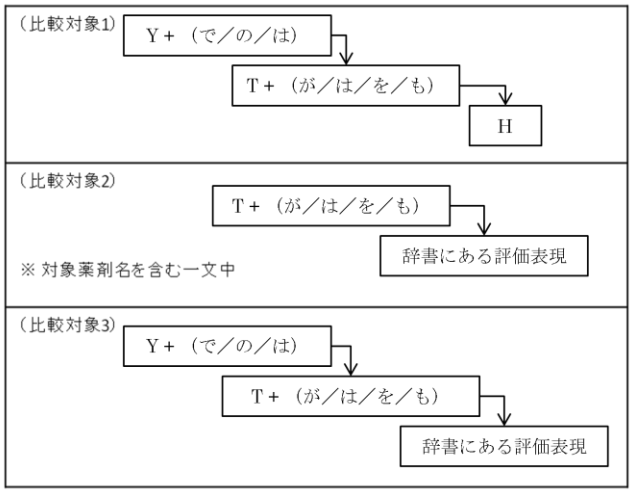
副作用	影響	おかげ	せい
使用	服用	投与	後
飲む	使う	増える	変える

文節を発見する．ここで，ひとつの長方形が一文節を，矢印はCaboChaを用いて求められる係り受け関係を示している．服用に関すること以外も多く書かれたブログ記事を対象としているため，抽出ノイズが減少するように予備調査から得られた服用情報の出現特徴を用いたいくつかの条件を付与する．２．１でタグ付けされた服用情報から，対象要素の文節はどのような助詞により評価要素存在文節へと連結しているかを調べたところ上位4件は「が(57.7%)」「は(12.7%)」「を(11.0%)」「に(3.9%)」であった．そこで，割合が高かった上位3件の助詞を用いて，評価要素の存在文節に最も近く，係る文節の最後尾が「が，は，を」であるとき，その文節を対象要素が存在する文節とする．また，服用表現を含む文節が，動作・状態の継続や経過に関する意味，原因・動機などの意味を持つ「で，から，より」，「が，ので，けれども」などの格助詞，接続助詞で終わるものに限定する．



Y:薬剤，TE:服用表現，T:対象，H:評価

図2　提案手法抽出パターン



Y:薬剤，T:対象，H:評価

図3　比較対象の抽出方法

２．４．三つ組生成

三つ組の対象・評価要素を含む文節から，助詞などの不必要な単語を除去する処理を行う．文節の先頭単語の品詞と同じ品詞の単語が後続していたとき，例えば共に名詞である「肝」と「機能」があったときそれらをつなげて「肝機能」として抽出する．「青あざ」のような先頭単語「青」が接頭詞の場合は，接頭詞とそれに後続する単語「あざ」をつなげ，後続単語「あざ」の品詞（名詞）を先頭単語の品詞と考え同様の処理を行う．また，要素の最後にくる単語に原形が存在する動詞の場合は原形とする．連体化の助詞「の」により他の文節から修飾されている場合は，その文節ごと要素の先頭へと追加し抽出する．

さらに，打ち消しの意味をもつ助動詞「ない」が文節中に奇数個存在した場合は，「ない」を付加した否定形で抽出する．なお，手がかり語存在文節の最後尾が逆接助詞であることや，「切る」「減量」といった手がかり語による意味の反転，「～ません」「～ませんでした」といった否定の表現は今回考慮していないが，今後極性判定を行う際にはそれらを考慮する必要がある．

３．評価実験

３．１．実験方法

TOBYO辞典よりランダムに取得した2,000スニペットから，各対応薬剤名を含む2,369種の文を用いる．人手で抽出した249組の〈薬剤・対象・評価〉からなる服用情報を正解データとする．

また，従来の意見抽出方法との比較を行うため，図3で示す3手法を比較対象とした．比較対象2，3で用いる辞書は，本研究の対象ドメインと一致する評価表現辞書が存在しないため，小林ら[7]の作成した一般的な評価表現辞書であるEVALDIC_ver1.0.1を利用する．

提案手法と，比較手法1～3により自動抽出された三つ組の服用情報としての適切さを，第一著者が判断する．評価には以下の式(1),(2),(3)を用いる．

$$\text{精度} = \frac{\text{正解抽出三つ組数} \times 100}{\text{抽出三つ組数}} \ [\%] \qquad (1)$$

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

$$\text{F値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}} \qquad (3)$$

表3　正解出力例

システム	入力文	出力三つ組
提案手法	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）
	反対にリウマトレックスを飲んで翌日からしばらく吐き気と倦怠感がけっこうつよく今、痛みが出てきていると「早く月曜日になって〜（月曜はリウマトレックスの服用日）」と思ったり	（リウマトレックス， 痛み， 出る）
	よかった頭痛については、プレドニンを投与しはじめて極度に視力が下がってきていてコンタクト	（プレドニン， 視力， 下がる）
比較対象1	”リフレックス服用後のムズムズ感が無くなった事を伝えたあとの先生からた	（リフレックス， ムズムズ感， 無くなる）
比較対象2	エンブレルを開始してまだ3回目なのですが、痛みが激減して喜んでいま	（エンブレル， 痛み， 激減）
比較対象3	私にとってタキソテールの脱毛率は結構高い	（タキソテール， 脱毛率， 高い）

表4　失敗出力例

システム	入力文	出力三つ組
提案手法	いままで甲状腺ホルモンが沢山でていたが、いくら間食・暴飲暴食しても体重が増えないはずだ!!メルカゾールのおかげで病気自体はよくなってきたものの私の体がどんど	（メルカゾール，私の体，どんど）
	ペガシス単独治療を終了して1年半以上を経過した	（ペガシス， 1年半以上， 経過）
	イレッサ隔日服用に!今日はS先生の外来日で肝臓関係の血液検査を診ていただきました	（イレッサ， !今日， 診る）
比較対象1	ただパキシルのほうが新薬で改良型なので吐き気の具合は改善されるかも	（パキシル， ほう， 改良型）
比較対象2	（左側目に傷がついていたせいで点眼薬を出してもらいました（ヒアレイン）次回の予約は3週間後です担当のセンセは9月30日付で異	（ヒアレイン， 予約， 異）
比較対象3	ずーーといい天気が続いていますここは、家の前は太平洋なので冬は穏やかな日が結構多いのです関節リウマ朝食後に普段は6錠だけど、今日は7錠プレドニゾロン2錠、ボルタレンSRカプ	（ボルタレン， 前， 太平）

３．２．結果とその考察

それぞれの抽出方法による出力結果のうち，正解例を表3に，失敗例を表4に示す．実験結果を表5に示す．

提案手法により47組の三つ組を抽出することができ，そのうち23組が服用情報として適切であったため，精度は48.9%，F値15.5という結果になった．また，正解データ数249より，再現率は9.2%であった．これについての原因は４．２．で考察する．比較対象1〜3により抽出された三つ組数はどれも提案手法より多いが，精度はすべて10%を下回っており，F値は提案手法が最も高い結果となった．

比較対象2では薬剤名と対象・効果の関係を考慮していないため，抽出三つ組数が多くなったと考えられる．また，服用情報の評価要素として評価表現辞書にある表現を用いることが適切でないことが，比較対象2，3の精度が上がらない原因だと考えられる．さらに比較対象1，3の結果から従来のように「薬剤→対象」「対象→評価」の係り受け関係のみでの服用情報抽出は闘病ブログに対し不十分であることがわかる．「薬剤＋服用表現→評価」「対象→評価」の係り受け関係から服用情報を抽出することが本手法の対象テキストには適していることが確認された．

表5　実験結果

	抽出三つ組数	正解三つ組数	精　度 (%)	再現率 (%)	F値
提案手法	47	23	48.9	9.2	15.5
比較対象1	121	8	6.6	3.2	4.3
比較対象2	1005	82	8.2	32.9	13.1
比較対象3	54	4	7.4	1.6	2.6

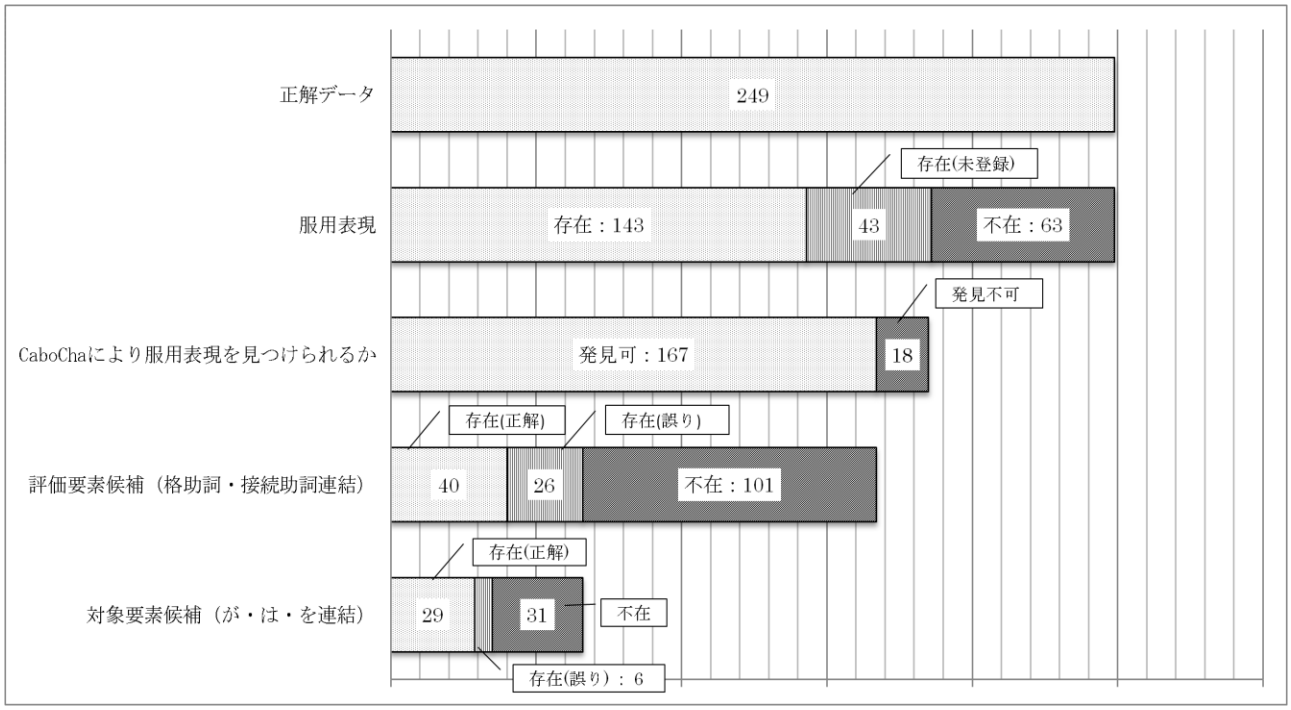


図4　正解データ抽出解析結果

しかしながら，比較対象1，3により抽出された正解三つ組はすべて提案手法では抽出することができないものであった．それらは服用表現が存在しない，パターンマッチングに当てはまらないなどの問題から提案手法では抽出できない．そのため，提案手法により三つ組を抽出できない場合には，それまでに抽出された要素からなる医療用評価表現と従来のパターンマッチングを用いて抽出を行うなど，両方の利点を生かした手法への改良が必要である．

４．誤りの解析

提案手法により抽出できたが適切ではないもの，人手で服用情報があると判断したが提案手法により抽出することができないものについてそれぞれ誤りの解析を行う．前者は精度向上，後者は再現率向上の足がかりとなる．

４．１．抽出誤り

三つ組すべての要素を抽出したが，適切でない場合についての考察を行う．

誤りと判断した理由として，要素が不十分なこと，意味的に適していないことが挙げられる．不十分な要素は，連体化助詞「の」以外の文節よって修飾されていたこと，もしくはスニペットを利用したため入力文が完全な一文ではないことが原因であり，要素あたりの情報として不足してしまっているものである．意味的に適さない要素とは，主に対象要素にみられ，日時や期間を要素としてしまったものである．今回処理時間などの問題からスニペットを対象としているが，これらを防ぐためにブログの全文検索を行うことや，スニペット記事に適切な助詞などを補完したのち抽出を行うことが必要である．本手法では要素の意味まで考慮した抽出を行っていないが，意味的に適さない要素かどうかの判断を単語の上位概念によるフィルタリングや機械学習を用いることで可能であると考えられる．

４．２．抽出不可

正解データ249文から，なぜ三つ組を抽出することができなかったのか考察を行う．

提案手法は服用情報の要素を，薬剤名，服用表現，評価要素，対象要素の順に抽出している．それぞれの抽出状況を表6に示す．表6より，服用表現が存在しないこと，もしくは登録されていないこと，服用表現後助詞の限定，対象要素助詞の限定により抽出不可能な原因であることがわかる．なかでも服用表現に後続する助詞を格助詞・接続助詞へと限定したことによる抽出もれが最も多い．そこで，格助詞・接続助詞の限定付けを行わずに，評価実験で用いたデータと同じ2369文から服用情報を抽出した．その結果，出力三つ組の数は110組となり，うち正解三つ組は32組存在した．抽出された三つ組数は約2.3倍に増加したが，正解数は1.4倍しか増加しておらず，ノイズ増加率が高い．精度よく抽出するには４．１でも述べたとおり，構文情報のみを用いて抽出するだけでは不十分であり，その要素の意味まで踏み込んだ解析を行うことが必要である．抽出する要素

候補を増やすという点では，助詞の条件付けを緩和することや，並立関係を考慮すること，抽出に用いる服用表現を増やすことが有効であると考えられる．

５．まとめと今後の課題

闘病ブログスニペットから三つ組で表現される薬剤の服用情報を，服用表現と係り受け関係を用いたパターンマッチングにより抽出する手法の提案を行った．本手法により，ドメインに依存するため作成コストのかかる評価表現辞書を使用せずに服用情報の抽出をすることが可能となった．従来のような商品レビューからの意見抽出に用いる方法では，闘病ブログスニペットからの服用情報抽出は10%未満であった．それに対し，提案手法では精度48.9%，F値15.5%と，既存の手法に比べ高い精度で闘病ブログスニペットから服用情報を抽出できることが確認された．

今後の課題としては，機械学習を用いた抽出要素の適正判定を行うことや，構文解析の誤りを減らすために並立助詞の利用や文の補完を行うことでF値を向上させる必要がある．また，抽出に用いる服用表現を類語辞典や学習から自動的に増加させることで，より多くの文に対する提案手法を用いた抽出の可能性が期待できる．さらに，提案手法により抽出不可能な場合は従来の手法を用いるなど，複数の抽出方法を組み合わせることでより効果的な抽出が可能となると考えている．

服用表現の抽出を高精度で行うことができるようになった後には，その情報を用いた極性判定を行う．極性判定結果と，ブログの時系列情報を利用し，ある薬剤の服用情報の変化を視覚的に表現することで医療支援となるシステムの構築を目指す予定である．

参考文献

[1] 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

[2] 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，2008

[3] 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

[4] TOBYO::日本の闘病記 32000
http://www.tobyo.jp/

[5] CaboCha: Yet Another Japanese Dependency Structure Analyzer
http://chasen.org/ taku/software/cabocha

[6] MeCab: Yet Another Part-of-Speech and Morphological Analyzer
http://mecab.sourceforge.net/

[7] 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.

連絡先

北海道大学情報科学研究科
E-mail:shihov_vo@media.eng.hokudai.ac.jp