文章構造に基づいたBlogからの体験情報抽出方法の提案

高野　太希† 　井上　潮‡

†東京電機大学工学研究科 〒101-8457 東京都千代田区神田錦町 2-2

E-mail: †09kmc24@ms.dendai.ac.jp, ‡inoue@c.dendai.ac.jp

あらまし　近年，Blogを使って個人から多様な情報が発信されている．発信された情報には多くの体験情報が含まれており，実際に見た対象物や，何かを体験した際の感想など，現実世界の事物に対する客観的あるいは主観的な情報が書かれている．こうした情報は，他のユーザに対しても有用であり，意思決定や問題解決などに役立てることができる．Blogは検索エンジンを使ってキーワード検索ができるが，体験情報を含むものだけを抽出することはできない．本稿では，Blogの書き手の一般的な傾向に基づいて体験情報を「体験を示す語」と「体験しなければ記述できない表現」の２つに分類し，それぞれに対して抽出条件を作成することによって体験情報を含むBlogのみを選別するとともに，体験情報が記述されている部分を抽出するための手法を提案する．

キーワード　情報抽出・情報要約

An Extraction Method of Experience Information from Blogs Based on Sentence Structure

Daiki TAKANO† and Ushio INOUE‡

†Tokyo Denki University 2-2 Kanda-Nishiki-cho, Chiyoda-ku, Tokyo, 101-8457 Japan

E-mail: †09kmc24@ms.dendai.ac.jp, ‡inoue@c.dendai.ac.jp

Keyword　information extraction, information summarization

1. はじめに

Blogは，著者の個人的な体験や日記，時事問題等，現実の世界での情報を記述し，公開している．これらの情報の中には多くの体験情報が含まれている．体験情報とは，実際に見た対象物や，何かを体験した際の感想など，客観的あるいは主観的な情報である．例えば,「○○タウンのイルミネーションが綺麗だった」や「○○祭り当日は多くの人で大混雑でした」のような情報である．このような体験情報には，イベントを体験した本人からの情報が書かれているため,主催者(ホームページ等)が提供していない情報が得られる.従って，イベントへの参加を検討している閲覧者にとって有用である．

Blogは複数のBlogエントリから構成されており，キーワード検索を行うことによって指定したキーワードに関するBlogエントリを検索することができる.しかし，総務省の「ブログの実態に関する調査研究の結果」によると，2008年1月の時点で国内のBlogエントリの総数は約1,700万件にも及んでいる．実際にイベント名を用いてキーワード検索を行うと，検索結果の件数が多く，さらに「未体験または予想の記述」や「宣伝や告知，広告」といった体験情報ではないものも多く含まれている．

そこで我々は,キーワード検索で得られたBlogエントリ集合の中から体験情報を自動的に抽出し，閲覧者の負担を軽減することを目標とした．

本稿では，Blogエントリに記述された日本語文章の構造を解析することにより,体験情報を含むBlogエントリを選別するとともに，有用な情報を判別するための手法を提案し，その有用性を検証する．

2. 先行研究

Blog解析に関して，多くの研究が行われている．本研究に近いものとしては，商品やサービスの評価情報の抽出の研究がある．鈴木ら[1]は，評価表現を抽出しつつ，発言全体が肯定的か否定的かの評価を判定する手法を提案しており，本提案同様，ユーザの意思決定や危機管理などの支援を目的としている．しかし，個人の評価を収集する点で経験の一部を担っているが，「美味しい」や「うるさい」など，明示的に評価を述べた記述から評判情報を抽出するという課題設定となっている．そのため,「○○があった」や「○○が披露されていた」のような現地での情報やトラブルといった個人の体験をすべて収集することはできない．

一方池田ら[2]は，ユーザの体験から得られる評判情報を抽出することを目的としており,「○○を買った」

や「○○を食べた」といった情報の抽出にも成功している．しかし，文章中での「対象物」と「評価表現」の距離によって判定する手法であるため「対象物」が記述されていない評価のみの文章からの抽出が困難である．また，「評価対象物」と「ユーザの求める対象物」の一致度が考慮されていないため，対象物の名称が完全に一致したものしか抽出できず，名称の書き方に揺らぎがあると抽出できないと考えられる．

3. 提案手法

キーワード検索によって得られた Blog エントリ集合から体験情報を抽出するために，Blog エントリ内の文章に対する構造解析によって有用性の自動判定を行う．本システムの処理の流れを図 1 に示す．

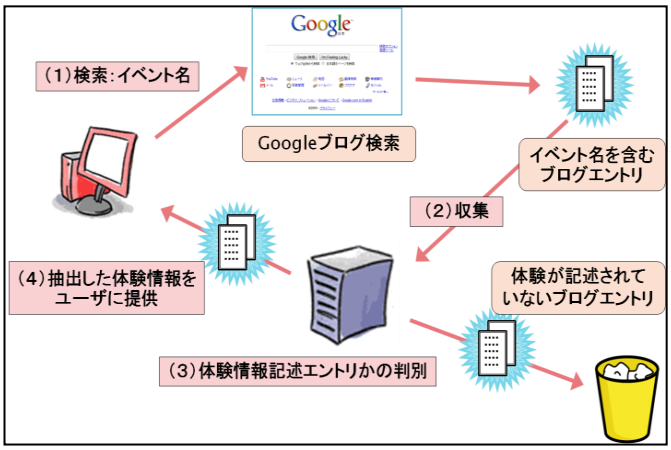

図 1　体験情報の選別と抽出システム

3.1. 選別する体験情報の定義

Blog エントリを，以下の 3 種類に分類する．

(1) イベントの体験情報を含む Blog エントリ
(2) イベント関連情報を含むが，体験情報を含まない Blog エントリ
(3) イベントとは関係のない Blog エントリ

イベントの体験情報とは，実際にイベントに行って得られた情報とする．具体的な体験情報としては，「大賑わいでした」や「繰り広げられていた」といったその場の状況を示す表現や「楽しめました」や「良かった」といった内容に対する感想などである．実際にイベントへは行っていないが，知っている情報（伝聞した情報）を記述したものや，イベントの開催日や行われる内容の告知は書き手が体験したものではないため体験情報ではない．

本提案では(1)のみを選別対象とする．表 1 に体験情報の例を示す．

表 1　本提案で扱う体験情報の例

	例文	判定
a)	昨日は○○祭りへ行ってきました。	○
b)	○○タウンのイルミネーションが綺麗だった。	○
c)	○○がとても面白かった。	○
d)	○○祭り当日は多くの人で大混雑でした。	○
e)	○○の歌や踊りが披露されていた。	○
f)	明日は○○祭りへ行きます。	×
g)	会場では○○が売られる。	×
h)	○○に感動する。	×

例文 a)は「見てきました」や「参加しました」のような書き手が行動したことついての記述であるため体験情報である．例文 b)，c)も「満足だった」や「楽しかった」のような書き手の体験から得られる感想であるため体験表現とする．例文 d)，e)は行動や感想ではないがその場の状況を示す情報であり，体験した本人にしか記述できない表現のため体験情報である．一方，例文 f)は書き手がこれから行動するという記述であるため体験情報ではない．例文 g)はその場の状況を示す記述であるが，書き手が体験したものかどうかは特定できないため体験情報ではない．例文 h)についても感想を表す記述であるが書き手が体験したものかどうかは特定できないため体験情報ではない．

3.2. Blog エントリに記述されている本文の抽出

本提案では，キーワード検索によって得られる検索結果の収集に GoogleAjaxSearchAPI[9]を使用している．しかし，GoogleAjaxSearchAPI では Blog エントリの要約のみしか収集することができないため，各 Blog エントリから記述されている本文のみを抽出しなければならない．Blog エントリ本文を抽出するために Perl モジュールである HTML::ExtractContent を使用する．goo ランキング[10]に記載されているイベント名上位 30 件から得られた Blog エントリのデータ合計 1740 件を使用して本文抽出精度を測った．平均精度を表 2 に示す．

表 2　本文抽出の平均精度

合計件数	合計成功数	合計失敗数	平均精度[%]
1740	1589	151	91.3

抽出に失敗したものは「You don't have permission to access」といった内容が格納されており，モジュールのアクセスを拒否された場合である．また，写真に対するコメントのみや、極端に文章が短いと誤抽出されてしまった．さらに特定のドメインの Blog(ameblo 等)に関しては失敗が多く見られた．精度結果として 9 割以上の Blog エントリの本文を抽出することができた

ため，HTML::ExtractContent を使用することにした．

3.3. 文章解析の処理の流れ

各 Blog エントリから抽出した本文に対する処理の流れを「先週、友達と博多どんたくを見に行きました。」という文章を例として図 2 に示す．なお，形態素解析には Mecab，係り受け解析には Cabocha を使用している．

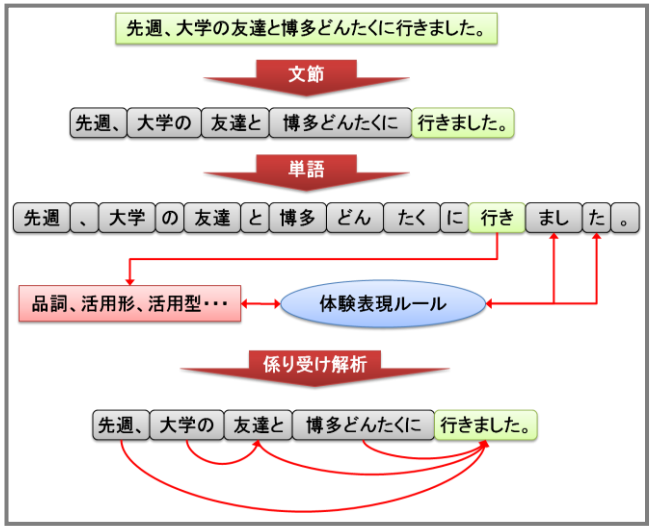

図 2　文章解析の処理の流れ

各文章に対して解析を行うために，Blog エントリに記述されている本文を文章単位に分割する．次に，分割した各文章に対し形態素解析及び係り受け解析を行う．解析結果を基に体験表現ルールと比較を行い，体験情報を抽出し，選別条件別の処理を行う．

3.4. 体験情報を含む Blog エントリの選別方法

予備調査として goo ランキングに記載されているイベント名 30 件(付録)のうち上位 15 件から Blog エントリ 1000 件を収集した．収集した Blog エントリの中で，イベントの体験情報が記述されていたものは 406 件である．イベント名が「実体験を示す語」(例：「行ってきました」，「見てきた」等)に係っている場合が 54.5[%]，「実体験を示す語」が記述されていないが「実際にイベントを体験しなければ記述できない表現」が記述されている場合が 36.3[%]であった．そこで，以下の 2 つの選別条件のどちらかに当てはまる場合にイベントの体験情報を含む Blog エントリとすることで約 9 割の体験情報を抽出することができると考えた．

3.4.1. 選別条件 A

抽出した本文に形態素解析と係り受け解析を行い，この情報から「実体験を示す語」の活用形を基に，係っている語を抽出する．「実体験を示す語」は 3.4 の予備調査において出現する頻度の高い語を選ぶ．抽出した語がキーワード検索で用いたイベント名であった場合，その Blog エントリに体験情報が記述されていると判定する．処理の流れを図 3 に示す．

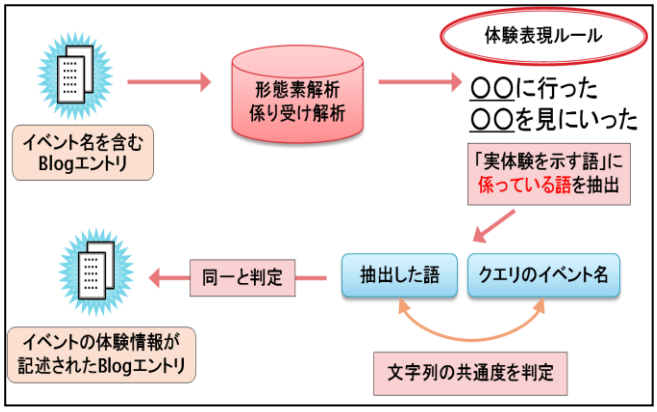

図 3　選別条件 A の処理の流れ

3.4.2. 選別条件 B

「実体験を示す語」が記述されていない場合，「実際にイベントを体験しなければ記述できない表現」が記述されているかを判定し，イベントの体験情報の有無判定を行う．

「実際にイベントを体験しなければ記述できない表現」は「楽しかった」等の著者自身が観察したことを表す表現が多い．これらの表現のうち予備調査から確実性の高い表現の体験表現ルールを作り，その体験表現ルールに適合している場合，Blog エントリに「実際にイベントを体験しなければ記述できない表現」が記述されていると判定する．また，その表現がブログ検索のクエリのイベント名に対するものであるかを判定するために，Blog エントリのタイトルにイベント名を含むものに限定する．処理の流れを図 4 に示す．

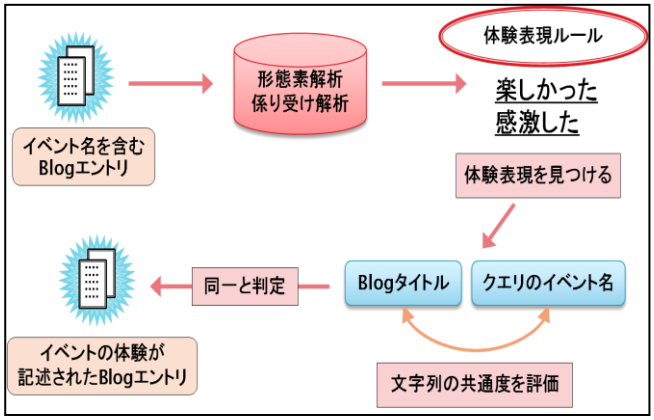

図 4　選別条件 B の処理の流れ

3.5. 各体験表現ルールの作成

図 1 における(3)の処理である「体験情報記述エントリかの選別」をするために，予備調査で「実体験を示

す語」に分類された文の集合に，形態素解析と係り受け解析を行い，形態素の出現頻度及び係り受け関係を基に表現の特徴を分析し，選別方法 A，選別方法 B について体験表現ルールを作成した．作成した体験表現ルールの代表的なものを表 3，表 4 に示す．予備調査において，選別方法 A で対象としている文の数は 215 であり，選別方法 B で対象としている文の数は 140 であった．対象となっている文であっても，体験情報ではない表現と重複する表現は除外したため，各体験表現ルールの割合の合計は約 9 割となっている．

選別条件 A の「実体験を示す語」として「行く」「見る」等の連用形，連用タ接続及び特定の名詞(サ変接続)を対象としている．対象としているサ変名詞は展示，感動，満足，びっくり，ビックリ，感心，興奮，感激，満喫，堪能，実感の 11 語である。

表 3　選別条件 A の体験表現ルール(一部)

品詞	活用形	語	例	割合[%]
動詞	五段・カ行促音便+連用タ接続	行く	行ってきました	57.5
動詞	一段+連用形	見る・訪れる・出かける	見てきました	23.1
動詞	五段・カ行促音便+連用形	行く	行きました	7.4
名詞+サ変接続	なし	参加など計 11 語	参加した	5.6

表 4　選別条件 B の体験表現ルール(一部)

体験表現ルール	表現の例	割合[%]
形容詞の連用形(連用タ接続)	「面白かった」「よかった」など	55.7
形容詞，形容動詞，副詞，サ変名詞+「でした」「だった」	「きれいでした」「満足だった」など	21.4
サ変名詞+「する」の連用形+「た」	「感動した」「堪能した」など	10.0
動詞，サ変名詞+接続助詞「て」+補助動詞「いた」	「盛り上がっていた」「賑わっていた」など	8.6
サ変名詞+動詞(未然形)+接続助詞「れ，られ」+接続助詞「て」+補助動詞「いた」	「展示されていた」「開催されていた」など	3.6
動詞(未然形)+接続助詞「れ，られ」+接続助詞「て」+補助動詞「いた」	「行われていた」「売られていた」など	0.7

3.6. 文字列の共通度判定

体験表現ルールによって判定された表現がクエリのイベント名に対するものであるとは限らない．そのため，抽出したイベント名候補と，クエリのイベント名の各文字列の共通度によって同一のイベントを示しているかの判定を行う．ユーザの求めるイベント名候補を「河津桜祭り」，抽出したイベント名候補を「河津の桜祭」として処理の流れを図 5 に示す．

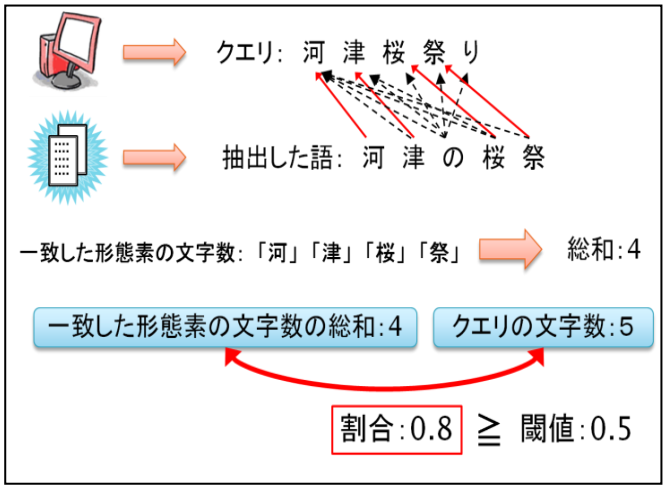

図 5　文字列の共通度による処理の流れ

3.6.1. 閾値決定実験

goo ランキングに記載されている上位 30 位のイベント名を用いて収集した Blog エントリ 2000 件に対し，文字列の共通度判定で使用する閾値を決めるための実験を行った．使用する閾値を 0.1 刻み(0.1～9.0 の間)に変化させて，抽出したイベント名候補とクエリのイベント名が同一であると正しく判断できる精度を求めた．正確性と網羅性の観点から評価を行うために F 値を使用した． R：出力のうちのクエリのイベント名の数，N：システムがイベント名とした数，C：すべてのイベント名候補の数として，適合率(precision)，再現率(recall)，F 値を次式で求めた．結果を表 5 に示す．

$$\text{適合率} = \frac{R}{N}, \quad \text{再現率} = \frac{R}{C}$$

また，適合率と再現率の調和平均である F 値(F-measure)は，

$$\text{F 値} = \frac{2*\text{precision}*\text{recall}}{\text{presicion}+\text{recall}}$$

によって求められ，この F 値が高ければ性能が良いことを意味する．

表 5　実験結果

閾値	適合率	再現率	F 値
0.1	0.34	0.96	0.48
0.2	0.55	0.95	0.68
0.3	0.79	0.92	0.82
0.4	0.82	0.84	0.79
0.5	0.95	0.82	0.85
0.6	0.96	0.64	0.71
0.7	0.98	0.56	0.69
0.8	1.0	0.38	0.58
0.9	1.0	0.37	0.56

閾値を上げるにつれ，適合率は高くなるが，表現が少しでも異なると判定漏れが生じるため再現率が低くなる．一方，閾値を低く設定すると表現が異なっていても同一と判定できるため再現率は高くなるが，間違った候補も判定されてしまうため適合率が下がる．イベント名の判定においては適合率と再現率のどちらも同程度重要であると考えられるため，F 値が最も良かった閾値 0.5 を使用することにした．

4. 評価

4.1. システムの出力結果

選別条件 A の出力例を図 6，選別条件 B の出力例を図 7 に示す．

検索するイベント名を入力してください：
博多どんたく
1956 : http://www.wieder.co.jp/2010/05/04/dontaku2010/
どんたくに行きました

1957 : http://fflc.exblog.jp/10543187/
博多どんたくに行ってまいりました

1973 : http://minkara.carview.co.jp/userid/302746/blog/18056944/
先週３日、大学の友達と博多どんたくを見に行きました

1973 : http://minkara.carview.co.jp/userid/302746/blog/18056944/
２１年福岡で育ったのに自分は博多どんたくを生で見に行ったことがありませんでした（汗）

1981 : http://haru-yoshi-yu.blog.so-net.ne.jp/2010-05-06
全国的に有名な「博多どんたく」ですが、実は、はーちゃんだけでなく、パパもママも初めて見ました

1999 : http://miwachan18.exblog.jp/13614442/
５月３日、博多どんたくに参加しました

図 6　選別条件 A の出力例

text	tag	phrase	keyword
…なるほどそうだったんですねぇ	そう	…なるほどそうだったんですねぇ	博多どんたく
九州内の複数の多彩な祭り・踊りが見れると、個人的に一番楽しみにしていたセッションでありますす	する	していた	博多どんたく
迫力の生歌・生演奏でした	演奏	生歌・生演奏でした	博多どんたく
沿道の僕らも日なたはつらかったですから	つらい	つらかったですから	博多どんたく
子どものひょっとこさんも、とても上手でした	上手	上手でした	博多どんたく
いやぁ…噂通りすごいパレードでした	パレード	パレードでした	博多どんたく
あらら～、でも、美しかったです	美しい	美しかったです	博多どんたく
パレードは大盛況だったのですが、気温はすごく高く暑かったです(^^;	暑い	暑かったです	博多どんたく
舞台が狭くて踊るのが難しかったけれど、お客さんが近くまでたくさんいてとても良い舞台でした	難しい	難しかったけれど、	博多どんたく
アクロスの舞台は、室内というのもあってか、お客さんのノリがいまいちだったのが残念でした	いまいち	いまいちだったのが	博多どんたく
アクロスの舞台は、室内というのもあってか、お客さんのノリがいまいちだったのが残念でした	残念	残念でした	博多どんたく

図 7　選別条件 B の出力例

4.2. 評価方法

選別条件 A，選別条件 B それぞれに対し，goo ランキングに記載されているイベント名 30 件のうち，予備調査で使わなかった 16 位から 20 位のイベント名に対して Blog エントリを各 64 件ずつ選択し，本文を取得できなかった記事やスパムブログ 34 件を除いた 286 件を用いて評価を行った．

4.3. 選別条件 A

平均精度を表 6，各イベントに対する精度を表 7 に示す．

表 6　選別条件 A の平均精度

適合率[%]	再現率[%]	F 値[%]
81.8	61.2	69.5

表 7　各イベント名に対する精度

イベント名	適合率[%]	再現率[%]	F 値[%]
高山祭	69.2	60.0	64.2
鶴岡八幡宮の流鏑馬	85.7	50.0	63.2
おわら風の盆	100.0	66.7	80.0
長岡まつり大花火大会	70.6	66.7	68.6
長崎くんち	83.3	62.5	71.4

適合率に関しては，81.8[%]の精度が得られた．しかし，「見る」と「みる」等の活用形や読み方では同じであっても使い方が異なるものは誤判定されていた．また，イベントに行ったこと(体験したこと)があっても体験した日時によって体験情報とは限らない．「○○に子供の頃に行った。」といった文章は「○○に」は「行った。」に係っているが情報としては古いため有用な情報とは言えず，誤判定された．このことから，「○○年○月に高山祭に行った」というような文章でも誤判定されてしまうと考えられる．しかし，Blog エントリの投稿された日時を検索内容に加えることで古い情報を除外することができると考えられる．

再現率に関しては適合率に比べ低い数値となった．これは，Blog エントリ本文に「実体験を示す語」がイベント名と一緒に記述されていないものがあったためである．さらに,使用した係り受け解析の制約により文章中に括弧が存在すると係り受けが正しく解析されないことや，Blog の著者による固有の表現によっても判定もれが生じた．再現率向上のためには，ルールの追加が必要である．

4.4. 選別条件 B

平均精度を表 8，各イベントに対する精度を表 9 に示す．

表 8　選別条件 B 平均精度

適合率[%]	再現率[%]	F 値[%]
70.8	63.4.	64.2

表 9　各イベント名に対する精度

イベント名	適合率 [%]	再現率 [%]	F 値 [%]
高山祭	65.0	76.6	68.3
鶴岡八幡宮の流鏑馬	85.9	58.3	69.5
おわら風の盆	67.0	67.0	66.7
長岡まつり大花火大会	80.0	54.2	58.3
長崎くんち	56.3	60.7	58.3

適合率に関しては 70.8[%]の精度が得られた．しかし,選別条件 A のような「語の限定」を一部のルールでのみ適用しているため,ルールが適用されない文章に関しては誤判定された．そのため，他の体験表現ルールに対しても出現頻度の高い語の調査を行い，適合率を向上させる必要がある．また,「よかった」という表現は「当日の天気はよかった」といった,天候に関する記述が多く誤判定された．

再現率に関しては 63.4[%]の精度が得られているが,実際にイベントに参加し,イベントで知り得たことに対する表現は「～はないようですね」,「～らしいです」など作成した体験表現ルールとは違う表現で書かれているため判定漏れが生じている．再現率の向上のためには「イベントに参加せずに伝聞した内容の記述」と「イベントに参加し,現地で見聞きした(知り得た)内容の記述」の違いを調べる必要がある．

5. まとめ

本稿では，体験情報はユーザの意思決定や問題解決に有用であると考え，Blog エントリに記述されている体験情報を自動的に抽出し，閲覧者の負担を軽減する手法を提案した．実際に投稿されていた Blog エントリを用いた評価の結果，約 65[%]前後の精度が得られた．実際に投稿されていた Blog エントリを用いて有用性の判定を行った結果「実体験を示す語」による抽出条件では約 70[%],「体験しなければ記述できない表現」による抽出では約 65[%]の精度となった．

今後の課題としては，まず、各評価結果で再現率が適合率よりも低くなっていることから体験表現ルールの追加による精度の向上が挙げられる．また，今回は体験情報であるか，体験情報でないかの二択で判定を行っていたが，有用性の強さについても考える必要がある．有用性が強い体験情報とは，多くの人が有用だと思う体験情報のことである．有用性が強い体験情報を優先して提示することで，閲覧者の負担がより軽減出来ると考えられる．

参　考　文　献

[1] 鈴木泰裕，高村大地，奥村学，“Weblog を対象とした評価表現抽出”，人工知能学会セマンティックウェブとオントロジー研究会，SIG-SWO-A401-02(2004).

[2] 池田佳代，田邊勝義，奥田英範，“体験表現を手がかりにした Blog の体験情報の抽出”，電子情報通信学会 第 18 回データ工学ワークショップ（DEWS2007）論文集 A8-1(2007).

[3] 土田正明，水口弘紀，久寿居大，“対象-属性-評価の 3 項同定による評判情報抽出”，言語処理学会第 13 年次大会(NLP2007)論文集，D2-3(2007).

[4] 倉島健，藤村孝，奥田英範，“大規模テキストからの経験マイニング”，電子情報通信学会 第 19 回データ工学ワークショップ（DEWS2008）論文集 A1-4(2008).

[5] 乾健太郎，原和夫，“経験マイニング:Web テキストからの個人の経験の抽出と分類”，言語処理学会 第 14 回 年次大会（NLP2008）論文集 C5-4（2008）.

[6] 水口弘紀，土田正明，久寿居大，“Weblog を対象にした評判情報分析システム eHyouban”，電子情報通信学会 第 19 回データ工学ワークショップ(DEWS2008) 論文集 I2-27(2008).

[7] 奥村学，南野朋之，藤木稔明，鈴木泰裕，“blog ページの自動収集と監視に基づくテキストマイニング”，人工知能学会セマンティックウェブとオントロジー研究会，SIG-ONT-A401-01(2004).

[8] 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)，75-82，2001-07-16.

[9] GoogleAjaxSearchAPI
http://code.google.com/intl/ja/apis/websearch/docs/reference.html#_intro_fonje

[10] goo ランキング
http://cache001.ranking.goo.ne.jp/crnk/ranking/051ki/yokoso_japan_festival/

付録

予備調査及び評価に使用したイベント名一覧

イベント名	イベント名
青森ねぶた祭り	高山祭
さっぽろ雪まつり	鶴岡八幡宮の流鏑馬
祇園祭	おわら風の盆
阿波おどり	長岡まつり大花火大会
大文字の送り火	長崎くんち
岸和田だんじり祭	明治神宮の初詣
仙台七夕まつり	御柱祭
精霊流し	コミックマーケット
博多どんたく	ロボコン
なまはげ	エイサーまつり
よさこい祭り	花笠まつり
博多祇園山笠	F1 日本 GP
東大寺お水取り	世界コスプレサミット
竿燈まつり	東京国際映画祭
東京ゲームショウ	成田山新勝寺の節分会