RD-003

口コミ情報からの目的情報抽出

Building a Database of Purpose for Action from Word-of-mouth on the Web

若木裕美† 有賀康顕† 中田康太†
Hiromi Wakaki Michiaki Ariga Kouta Nakata
藤井寛子† 住田一男† 鈴木優†
Hiroko Fujii Kazuo Sumita Masaru Suzuki

1 はじめに

従来の情報検索技術では，キーワード検索が主流であり直接的なキーワードが思い浮かばない場合には，ユーザ自身が工夫をして様々なページを閲覧し，有効なキーワードや条件を決めながら所望の結果を探す必要がある．しかし，最初から探したいものが決まっているとは限らず，例えば服飾品，レストラン，不動産，宿などの検索では，大まかな要求と対象とをすり合わせながら所望の結果を選択していくことになる．このような場合には，目的やジャンル程度の大まかな要求からシステムとのインタラクションに基づいてニーズを具体化しながら必要な情報に近接していける検索が必要である．例えば，具体的な地名や店名など (例：「大洗海岸」) をユーザは思いつかないことがあるため，目的（例：「海水浴」「子供を遊ばせたい」）を入力して目的に関連するものを検索できることが必要である．また，目的も複合的であることがあり，例えば，直接表現される目的（例：「海水浴」＝泳ぎたい）の他に，暗黙的な目的（例：「友達と騒ぎたい」「バーベキューがしたい」「焼けたい」「花火がしたい」）が隠されている場合があり，暗黙的な目的まで考慮した検索結果の提示ができると良い．そのためには，事前に目的についての情報を整理してデータベース化して直接目的から検索できること，さらには目的間の関係性を基にユーザの隠れた目的を把握する必要がある．

一方、インターネットが広く生活に浸透し，一般のユーザが自身の経験や見聞を基に様々な情報を発信するようになった．その結果，『価格.com[1]』や『Amazon[2]』や『じゃらん[3]』のサイトに代表されるように，商品やサービスの提供者側からの情報ではなくユーザ側から見た生の声を横断的に閲覧できるシステムが一般的になり，沢山の口コミ情報が簡単にまとめて見られるようになった．近年では，このような口コミ情報を有効利用するために，評判分析という分野で様々な解析・抽出・分類手法が研究されている [10]．さらに，2007 年頃からは評判を分析するだけでなく，個人の経験を広く Web 文書集合から抽出しデータベース化することを目指した経験マイニング [15][4] というアプローチも出てきた．本稿で提案する目的データベースは，個人の経験を基にその行動を起こす目的に着目し，データベース化することを目指したものである．

そこで，本稿では旅行情報を例題としてデータの分析を行い，目的データの定義を行った．目的データを「対象」「行動」「理由」の３つ組からなる情報と定義した．例えば，**「清水寺は紅葉で有名です．」**という文があった場合，**“ユーザが清水寺に行くのは，紅葉が有名だから”** といえ，目的データとして，

- 『対象』は「清水寺」
- 『行動』は「行く」
- 『理由』は「紅葉が有名」

のように記述する．さらに「理由」のタイプを５つに分類し，旅行口コミサイトの清水寺の口コミテキストに対して人手で目的データ抽出を試して妥当性を確認した．その結果「理由」の分類のうち最も表現のバリエーションが多かった「ポジティブな表現」を対象に，既存の辞書である極性評価辞書の表現を元に Google N-gram[11] を利用して極性判定を行い辞書を拡張する方法を提案する．

2 関連研究

評判情報抽出の分野では，対象・属性・評価という３つ組を抽出することが主な課題とされている．立石らは対象・属性・評価に関する共起パターンを利用して，属性表現と評価表現をブートストラップ的に抽出する手法を提案している [18]．杉木らは自然言語クエリによる評判情報に関する情報検索を行うため，抽出タスクで評価視点・評価値を抽出し，また検索タスクで検索対象・評価視点・評価値を抽出している [14]．倉島らは，非構造データであるブログデータを経験という観点で構造化することにより経験情報を検索可能にするため，状況（時間，空間）・行動 (動作，対象)・主観 (評価，感情) という人間の行動を軸とした経験そのものの情報抽出を行っている [15]．一方，我々は対象・動作・理由という３つ組により行動の目的を記述し，口コミデータから目的に関する情報抽出を行うことにより検索者の意図や目的に基づく検索を目指す．

情報検索の分野では意図理解のためにクエリ分類技術がある．クエリ分類（Query Classification）では入力された短いクエリが対応する話題カテゴリへ分類を行うことで検索精度向上を目指すものである [8][3][4]．クエリ分類で扱う意図とは，短いクエリに暗黙的に含まれる ‘話題’ であり，これを明示的に検索に利用することで文書検索の精度の向上を目指す．一方，我々の扱う目的情報では，検索者の行動（例：観光行動）の目的そのものをクエリとして検索可能にすることを目指し，クエリ分類で扱う意図よりもより細かい粒度を扱う．

目的データの項目の１つである理由は，評価表現抽出（辞書構築）と関連が深い．評価表現抽出には，WordNet[1] の隣接情報を利用してエントリの極性を判断する手法 [5]，コーパス中の特定の品詞の並びを利用して形容詞の極性を判断する手法 [2]，肯定否定の極性を持つ典型的な種表現と共起する比率に従って語彙の極性判定をする手法 [9]，ある評価表現の周辺文脈に着目して逆説表現がなければ同一の極性を持つと仮定して種表現からブートストラップ的に収集する手法 [17] などがある．また最近では，構文情報を利用して評価表現とともに属性や因果関係を抽出する手法が出てきた．高野らは，コーパスの構文解析結果に対して因果関係を持つパターンを手かがりと

†(株) 東芝 研究開発センター 知識メディアラボラトリー

[1] http://kakaku.com/

[2] http://www.amazon.co.jp/

[3] http://www.jalan.net/

[4] クエリ分類の研究が加速した要因のひとつとして挙げられるのは 2005 年に開催された KDD cup 2005 であり，約 80 万件の検索クエリを 67 のカテゴリに分類することをタスクとしていた．

して用意し，評価要因の抽出処理と評価表現の抽出処理を繰り返すことで，評価要因と評価表現をブートストラップ的に抽出する手法を提案している [12]．Qui らは，評価表現と製品の特徴情報（製品の構成要素や属性）の片方または両方から記述された幾つかの依存構造パターンを用意し，評価表現と特徴の両方を収集していく手法を提案している [6][7]．本稿では，目的データの項目の 1 つである理由として最も多かった「ポジティブな評価」の語の判定を細かい表現にまで適用可能にするため，すでにわかっている「ポジティブな評価語」に対して特定の構文パターンを構成する語を Google N-gram[11] から抽出することにより対象となる口コミデータ向けに語彙拡張する．

3 目的データ

3.1 旅行情報サイトの選定

旅行情報を題材として目的データの収集を行うため，適切な情報資源や文章について検討した．Web 上を対象に,「旅行ブログサイト (例：4travel[5], 旅スケ[6])」,「観光協会公式サイト (例：那覇ナビ[7])」,「旅行ガイドのサイト (例：All About の国内旅行ガイド[8])」，の 3 タイプの記載内容や量を比較した.「旅行ブログ」は自身の経験から人へのお勧めが,「観光協会のサイト」は年間のメインイベントなどを中心に有名スポットの紹介が事実記載的に,「旅行ガイドのサイト」は一般的な評価でお勧め情報が記載されている傾向があった．また,「観光協会のサイト」は質・量・記載形式がサイトごとにばらばらで収集が難しく,「観光ガイドのサイト」は記載量が少なく全国各地の情報を集めるのが難しい．そこで,「旅行ブログサイト」を対象にした．また最も口コミ数が多かった旅スケを対象として，記載された日本全国 47 都道府県の口コミデータ（計 40645 記事）をダウンロードし，これを以降の解析に利用することにした．

3.2 目的データの定義

ユーザの目的から直接検索可能にするため，事前に目的についての情報を整理してデータベース化したい．特に，ここでは旅行を例題として扱うため旅行の目的データベースを試作する．このため，目的データはどのような形式で記述されると良いかを上記の口コミ情報を見ながら検討した．そして，旅行を例題とした目的データ形式として，(1) 場所や物などの対象，(2) 旅行における行動，(3) その行動を取る理由，の 3 つ組を目的であると定義した．また，行動の目的となる理由にはどのような情報があるかを，口コミ記事のうち清水寺やその他京都の内容を中心に分析した結果，５タイプに分類できた．これらをまとめて次のように目的データを定義した．

目的データの定義

目的データとは、『対象』・『理由』・『行動』からなり、省略情報を補完した文から 3 つ組が取れるときであり、
(1) **行動**は、旅行目的の場合には旅行の行動
　(例：行く、楽しむ)、
(2) **対象**は、行動の対象となる 場所・物・体験
　(例：清水寺、夜景、キャンプ)、
(3) **理由**は、対象が目的になりうる理由で、
　(a) 特異性　(例：唯一、世界三大〜、最古)
　(b) 場所から連想されるイメージ
　　(例：京都っぽい、〜といえば、〜に限る)
　(c) 特徴的　(例：有名、名所、定番)
　(d) ポジティブな評価　(例：美しい、綺麗)
　(e) 対象者が限定される
　　(例：子供向け、〜が好きな人には)
のいずれかに該当する場合．

ただし，個人的な感想や，単なる事実の記載は含まれない。また，必須項目ではないが，場所，時期，その他の条件を補足情報として持つことができる．上の定義を使って文から目的データを抽出した場合，次のように表現される．

例

＜文＞「秋になると，清水寺は紅葉が素晴らしいことでも有名です。」
＜目的データ＞
(1) **行動**：　行く
(2) **対象**：　清水寺
(3) **理由**：　(代表) 有名,
　(詳細) 紅葉が素晴らしいことで有名
(*) 補足：　(時期) 秋

ここで，必須である対象・行動・理由（代表）の 3 つ組が取れるとき目的データであるとする．旅行を題材とした場合には ‘行動’ には旅行の行動（例：行く，楽しむ）が主に入ると考えられる．また，‘対象’ には行動の対象となる場所・物・体験（例：清水寺，夜景，キャンプなど）が該当する．なお，文単位では省略される情報もあるため，省略情報を補完した文から 3 つ組が取れる場合とする．

理由については，５タイプのいずれかに該当するものを理由として認めることにする．『特異性』とは，他にはない特徴があり非常に知られていることを示す表現であり，例えば「世界遺産」「日本三大夜景」等の表現である．『場所から連想されるイメージ』とは,「京都っぽい」「奈良といえば鹿」のように，暗黙に連想されたり形容されたりするイメージを指す表現である．『特徴的』とは，一般に知られている特徴を持つことを示す表現であり，例えば「有名」「名所」「定番」などである．『ポジティブな評価』とは，ポジティブな印象を伝えようとして用いる形容表現であり,「美しい」「綺麗」などである．『対象者の限定』とは,「子供向け」「カップル用」「昼食向け」などどんなタイプの人・どんな目的の人向けのかを記述した表現である．ただし，単なる事実や個人の体験は理由に含まれないものとした．また，例えば「ライトアップ」など，暗に綺麗 (= ポジティブな評価) なイメージがわくだけでは，理由とはしないこととする．

長いフレーズで記述された理由については，代表の理由と，それを説明する詳細の理由に分けることが必要と考えられる．また，付属する情報として，場所，時期，その他の条件も合わせて保持することができるものする．

3.3 目的データ定義の妥当性確認

本定義の妥当性確認のため，実際の口コミ投稿記事に対して目的情報が取得できるかをハンドシミュレーションし，同一記事での複数人間での一致度を測定した．

旅行目的データの定義に従い，清水寺の口コミ情報 60 件を利用して口コミ情報から目的データを抽出可能かを確認した．全 60 件の口コミ情報を 30 件ずつ２セットに分け，各セットを３名のメンバが担当して人手で目的データを生成，同一の目的データが抽出されるかを調査した．なお，テキスト中に省略箇所がある場合は，人手で補完してから目的データを抽出するようにした．各文ごとの目的データが抽出できる/できないの判断についての一致度は，0.79 となった[9]．さらに，生成された目的データの一致度は，0.57 となった[10]．抽出できる/できないの判断では概ね一致しており一般性がある定義といえる．しかし，口コミ情報が長く理由としてどこを中心に挙げるかが人によって分かれたり，補足情報と対象のいずれにも地名が入りうる曖昧性があるため何を中心に捉えるかでの不一致があった．

分析の結果，複文等の場合にテキストの途中で話題が変わる場合に目的をとるかで判断が分かれた．また，前提条件によって評価の分かれる「混んでいる（特異性）」や，ポジティ

[5] http://4travel.jp/domestic/
[6] http://tabisuke.arukikata.co.jp/domestic/ “地球の歩き方” が提供している口コミサイト
[7] http://www.naha-navi.or.jp/
[8] http://allabout.co.jp/domestic/
[9] ２人以上が目的データを抽出したか、あるいは３名全員が目的データを抽出できないと判断した場合，判定が一致したとみなす．
[10] ２名がほぼ同じ目的データを生成した場合，一致したとみなす

ブな文脈で使われることが多そうな「〜できる（ポジティブな評価）」が理由になるかどうかで判断が分かれた．このため，複文等の場合でも目的となる箇所が含まれていればその部分を抽出することにした．また前提条件によって評価が分かれるものは理由としないことにした．その結果，清水寺の口コミ情報 60 件 (307 文) からは，107 個の目的データ (99 文) が得られ，これを清水寺の目的データの正解セットとした．

なお，表 2 に清水寺の口コミ情報から人手で抽出した目的データ 107 個に含まれた理由タイプ別分類を示した．この結果から，(d) ポジティブな評価の語が半数以上を占め最も多く，(c) 特徴的な語が次に多いことが分かる．

4　目的データの収集のための理由抽出

4.1　なぜ理由の抽出を最初に行うか？

本章では，前節の目的データの定義に従い口コミ投稿記事から目的データを抽出する．口コミ投稿の各記事ではタイトルやタグが付与されたり，複数の文から構成されているため，必ずしも一文内に必要な情報が記述されているとは限らず，目的データの定義の 3 つ組のうち，『対象』や『補足情報』は省略されている場合がある．また，旅行を題材とした場合には『行動』は自明であるため，『行動』は記載されていない場合があり，最終的に『対象』や『行動』の省略を補う必要がある．一方，『理由』は目的データの定義の 3 つのうち省略され得ない箇所であるため，まずは『理由』の有無を見つけることが必要であると考えた．なお，本稿では理由抽出処理を行うところまでを実験する．

4.2　理由のタイプごとの語彙獲得方法・判定方法

前章では，『理由』には次の 5 タイプがあると分析した．『理由』抽出ではタイプごとに特性が大きく異なるため，各タイプ判定に必要な語彙またはパターンを用意し，各文でいずれかに一致する箇所を理由箇所として判定する．

理由の 5 タイプ

(a)　特異性 (例：世界遺産，日本三大夜景）
(b)　場所から連想されるイメージ
　（例：京都っぽい）
(c)　特徴的（例：有名、名所、定番）
(d)　ポジティブな評価（例：美しい，綺麗）
(e)　対象者の限定（例：子供向け）

(a) 特異性

『特異性』とは，他にはない特徴があり非常に知られていることを示す表現であり，主に (a-i)『数詞を伴う表現』、(a-ii)『登録名』の 2 つが考えられる．数詞を伴う表現としては，「日本三大稲荷」「日本三大夜景」のような一般的に評されるトップ 3 などを指す．登録名としては，「世界遺産」「国宝」のように登録機関によって決められるものを指す．

そこで，(a-i) 数詞を伴う表現は，「地名＋数字＋接頭詞＋名詞」のパターンに一致した表現を理由箇所として抽出する．また，(a-ii) 登録名は，収集した口コミ投稿記事中から「〜に指定」という表現に係る語（複合語）を収集し，辞書として用意した．理由判定時には，辞書にある語との一致で判断する．

(b) 場所から連想されるイメージ

『場所から連想されるイメージ』とは，「京都っぽい」「奈良といえば鹿」のように，暗黙に連想されたり形容されたりするイメージを指す表現である．一般にフレーズで表現されることが想定され自動収集は簡単ではない．そこで，本稿ではまず「〜に限る」「〜といえば、…だ」という 2 つの表現をパターンとして用意し，本パターンに合致した場合を理由と判定した．

[11] http://cl.naist.jp/ inui/research/EM/sentiment-lexicon.html

(c) 特徴的

『特徴的』とは，一般によく知られている特徴を持つことを示す表現であり，例えば「有名」「名所」「定番」などである．後述する極性評価辞書にも幾つか含まれたことから，(d) ポジティブな評価の語として収集できる可能性が高い．そこで，本稿では (d) に含まれるものとして扱う．

(d) ポジティブな評価

『ポジティブな評価』とは，ポジティブな印象を伝えようとして用いる形容表現であり，「美しい」「綺麗」などである．NAIST が提供している極性評価辞書[11]では，用言編 [13] と名詞編 [16] からなる辞書がある．用言編は，用言を中心に収集した評価表現約５千件のリストを一部改編し人手で評価極性情報を付与したデータであり「ポジティブ」「ネガティブ」「客観的」「主観的」の４分類の情報が付与されている．また，名詞編は，評価極性を持つ（複合）名詞，約 8 千 5 百表現に対して評価極性情報を付与し人手によるチェック済みのデータで，主観・客観表現についてポジティブ・ネガティブ・ニュートラルを付与されている．いずれも，辞書は単語と用法のセットで記述され，ある語がある用法で使われた場合に「ポジティブ」となると判断できる．

しかし，極性評価辞書は一般的な辞書であり，口コミ情報など多岐に渡る表現には対応しきれない．このため，判定対象である口コミデータ中の表現に対しても，ポジティブな語であれば収集して辞書に追加しておく必要がある．さらに，表 2 の結果から (d) ポジティブな評価の語が最も多く，また (c) と (d) が同一視される場合には全体の約 80% になることからも幅広い語が必要になるといえる．そこで，対象文書の表現に合わせた新しい辞書の拡張方法を提案する．本手法ついては次章で詳説する．

(e) 対象者の限定

『対象者の限定』とは，「子供向け」「カップル用」「昼食向け」などどんなタイプの人・どんな目的の人向けのかを記述した表現である．収集した口コミ投稿記事中から、「〜向け」「〜向き」という表現に係る語を集め、「向け」「向き」除いた語（複合語）を集め辞書とした．ただし，明らかに方角を示す語 (例：東，西，など) は人手で除いた．理由判定時には，文中で「〜向け」「〜向き」「〜用」「〜にお勧め」のいずれかのパターンの前にこれらの語が利用された場合にのみ理由として判定する．

5　ポジティブ評価語の語彙獲得

5.1　目的

『ポジティブな評価』とは，ポジティブな印象を伝えようとして用いる形容表現であり，「美しい」「綺麗」などである．NAIST が提供している極性評価辞書に存在する語については，ポジティブかネガティブか中立かが判断できるが，口コミ情報に含まれる多種多様な表現をカバーすることはできない．そこで，対象文書である口コミデータ中にごく少数回しか現れない表現であっても極性判定できるように Web データを利用した極性判定手法を新たに提案する．

表 1: 対象品詞と Cabocha による抽出パターン

	対象品詞	Cabocha を使った抽出パターン
(i)	形容詞	「形容詞」の単語
(ii)	形容動詞	「名詞 [形容動詞語幹]」＋「助動詞 (だ)」で終わる文節
(iii)	補助形容詞	「形容詞非自立 (やすい/がたい/づらい/にくい)」で終わる文節
(iv)	複合述語	「助動詞 (らしい/的)」で終わる文節

表 2: 清水寺の口コミ情報から人手で抽出した目的データの理由タイプ別分類

理由タイプ	抽出タイプ割合 (タイプ別抽出数/全抽出理由数)	具体例
特異性	7.5% (9/120)	世界遺産，日本十大名水
場所から連想されるイメージ	8.3%(10/120)	京都といえば，日本らしい
特徴的	25%(30/120)	有名，定番，人気，名所
ポジティブな評価	56%(69/120)	美しい，風情がある，最高
対象者の限定	1.7%(2/120)	カップル，婚活中

表 3: 利用した構文パターン

	構文「候補語 X ＋(助詞)＋極性語 Y」	意味	例
(I)	形容 (動) 詞 [連体] ＋ (なし) ＋ 形容 (動) 詞 [連体]	並列	美しい綺麗な
(II)	形容 (動) 詞 [連体] ＋ (なし)＋ 形容 (動) 詞 [名詞化]	形容	美しい綺麗さ
(III)	形容 (動) 詞 [連用] ＋ (なし) ＋ 形容 (動) 詞 [連体]	並列	美しく綺麗な
(IV)	形容 (動) 詞 [連用] ＋ て/で ＋ 形容 (動) 詞 [連体]	並列/理由	美しくて綺麗な
(V)	形容 (動) 詞 [連用] ＋ て/で ＋ サ変名詞	理由	美しくて感動

表 4: 日本全国 47 都道府県の口コミデータ（計 40645 記事，252692 文）から抽出された 98776 の理由について，理由タイプごとの抽出数．さらに，各タイプごとの理由箇所・文の例

理由タイプ	辞書・パターン	抽出数	(例) 理由箇所	(例) 元の文
(a-i)	希少性 1	174	日本三大庭園	広大な敷地がある**日本三大庭園**の水戸の偕楽園。
(a-ii)	希少性 2	2695	世界遺産	京都の**世界遺産**の 1 つ醍醐寺。
(b)	イメージ	359	高知といえば	高知**といえば**桂浜の坂本竜馬像ですよね。
(d)	極性辞書	63931	綺麗	また枯山水の日本庭園も非常に**綺麗**です。
	追加辞書	30804	幻想的	入り口を入ってすぐのイルカの水槽が、**幻想的**です。
(e)	対象情報	813	子供向け	この公園が大好きで、**子供向け**の遊具がありあす。

表 5: 清水寺の口コミ情報から自動で抽出した理由タイプ別分類

理由タイプ	辞書パターン	抽出タイプ割合 (タイプ別抽出数/全抽出理由数)
(a) 特異性	希少性 1	1% (1/93)
	希少性 2	4.3% (4/93)
(b) イメージ	イメージ	2.2% (2/93)
(d) ポジティブな評価 (+(b) イメージ +(c) 特徴的)		
	極性辞書	74% (69/93)
	追加辞書	17% (16/93)
(e) 対象者の限定	対象情報	1% (1/93)

5.2 手法の概要

口コミに現れるような多岐に渡る表現の極性判定を行うために，何らかのパターンの頻度を基に計算する場合には対象とする口コミデータよりもはるかに多くの文章を収集し解析する必要がある．しかし，そのような規模での文書収集は容易ではない．そこで，Google により公開された Google N-gram[11] の頻度情報を用いた手法を考案した．Google N-gram とは，Google がクローリングした Web ページ約 200 億文（約 2550 億単語）の日本語データから作成した単語 n-gram データ (1〜7 gram) を 2007 年 11 月に公開したものである[12]．通常 N-gram データがあっても，文全体があるわけではないので構文パターンの利用は難しいが，ここでは特定の構文を想定することで，N-gram データの範囲のテキストでその係り受け関係が推測できる部分を利用して，単語のポジティブさを測定する点に特徴がある．

5.3 手順

Step1: 形容詞・形容動詞の収集 まず，対象とする口コミ投稿記事の各文を日本語係り受け解析器 Cabocha [13] で解析し，形容詞・形容動詞を構成する文節を収集する．また，形容詞に似た表現として補助形容詞と複合述語も対象として収集する．特に，補助形容詞としては動詞の後に付いてある意味を付け加え形容詞と似た活用をする接辞として知られる「V+やすい」「V+がたい」「V+づらい」「V+にくい」を，複合述語としては形容詞的活用をする接辞で知られる「N-らしい」「N-的」を対象とした．なお，Cabocha では形容動詞という品詞はなく,「名詞 [形容動詞語幹]」+「助動詞 (だ)」で表されるため，本パターンに該当する場合に形容動詞と扱った．

実際の計算では，対象文書として旅スケの全国の口コミ投稿記事を利用した．また対象品詞と Cabocha を使った場合の抽出パターンは，表 1 にまとめた．

Step2: 構文パターンの生成 ここでは，短い構文で順接の意味になるパターンを生成し，辞書の拡張に利用する．まず，Step1 で用意した語をポジティブな語かを判定する候補語 (これを X とする) とする．また，X の後ろに順接の意味で形容詞・形容動詞 (これを Y とする) が続くような構文パターンとして，表 3 の 5 通りを用意した．各構文は，並列，形容，理由などの意味的関係を持つため順接になる構文であり，X と Y が同じ極性を持つ語が並びやすい構文である．これを用いて，Y には極性評価辞書でポジティブな語として用意されている単語をあてはめ，Step3 で Google N-gram の頻度から充分な数の Y を持つ X はポジティブらしいと判断する．なお，実際の文ではこの構文のさらに後ろに否定が続く可能性もあるが，複数の表現で合計を取るため否定形に続く語の影響は限定的と考えられる．

実際の計算では，まず極性語 Y として極性評価辞書から形容詞・形容動詞・サ変名詞を抜き出し，形容詞・形容動詞は，連体形と名詞化の形で，サ変名詞はそのままの形で用意しておく．また，Step1 で用意した語を候補語 X として，構文に合うように連体形，連用形の活用形を用意しておく．さらに Step3 では，終止形で同一語を集計するため，候補語 X の終止形も用意しておく．

Step3: Google N-gram を用いた極性推定 ここでは，Step2 で用意した候補語 X と極性語 Y から構成される表 3 構文 (I)〜(V) に合致する N-gram の頻度を Google N-gram から取得し極性を判定する．ここで，候補語 X_b(候補語 X の終止形を X_b とする) のポジティブさを次の式で定義した．

$$R_{positive}(X_b) = \frac{N_p(X_b)}{N_p(X_b)+N_n(X_b)} \quad (1)$$

ただし，極性語 Y のうちポジティブを Y_p，ネガティブを Y_n と記載する[14]．また，候補語 X と極性語 Y が構文パターン S の形になるものが Google N-gram 中に存在するかどうかを次の関数で表すことにする．

$$G(x,y,S) = \begin{cases} true, & \text{存在するとき} \\ false, & \text{存在しないとき} \end{cases} \quad (2)$$

さらに，

$$N_p(X_b) = |\{y \in Y_p | G(x,y,S) \wedge x \in \{X | base(X) = X_b\}\}| \quad (3)$$

とおいた．ただし，$base(X)$ は X の終止形を求める関数とする．なお，辞書拡張時には $R_{positive}$ の値が閾値以上の場合に，ポジティブな語であると判定し辞書に追加する．

実際の計算では，Google N-gram の 2〜6-gram 中[15]で，構文 (I)〜(V) の前半が候補語 X に存在する語であり，後半が極性語 Y に存在する語である場合を残す．そして，同一終止形 X_b の候補語 X 全てに関して，後ろに続く極性語 Y の異なり数に対するそのうちのポジティブな語の異なり数の割合を測定するのが式 (1) である．なお，ネガティブさを測る場合は同様にネガティブな語の割合とする．

5.4 実験

前節で定義したポジティブ評価尺度の式 (1) によりポジティブな語を収集する．極性評価辞書に登録されている語も含まれるため，収集結果の語の中で極性評価辞書にポジティブ，ネガティブ，中立のいずれかに存在した語を利用して精度・再現率を測定し，最も F 値の高くなる閾値を利用する．閾値と精度・再現率・F 値の変化を図 1 にまとめた．この結果から，最も F 値の高い 0.7 を閾値として利用した．その結果，旅スケから得られた全ての候補語から新たに収集できたポジティブ語は 921 語で，例えば「豪華絢爛だ」「緑豊かだ」「香ばしい」「京都っぽい」といった語が集まった．一方，同様にしてネガティブ語 (閾値 0.3) は，新たに収集できた語は 565 語であり，例えば「陰気だ」「わかりづらい」「閉鎖的だ」「不親切だ」といった語が集まった．収集語数は表 6 にまとめた．

ポジティブ語として収集された語彙の例
おしゃれだ, おちゃめだ, おめでたい, おもろい, お上品だ, お手ごろ価格だ, お手軽だ, お洒落だ, お花畑みたいだ, お買い求めやすい, かぐわしい, かっこいい, かっこよい, よさげだ, アクセスしやすい, アットホームだ, イタリアっぽい, エキゾチックだ, オトナっぽい, コミカルだ, コンパクトだ, サービス精神旺盛だ, ハイテクだ, バラエティー豊かだ, ルンルンだ, ロマンチックだ, レトロっぽい, 上品だ, 乙だ, 人なつこい, 伝統的だ, 住みやすい, 優美だ, 初々しい, 和モダンだ, 壮観だ, 大人気だ, 容姿端麗だ, 宿泊可能だ, 絢爛豪華だ, 隠れ家的だ, 気持ちよさそうだ, 小綺麗だ, 優しげだ, 緑豊かだ, 香ばしい, 京都っぽい, 感慨深い, 愉快だ, 情緒的だ, 歩きやすい, 流暢だ, 清涼だ, 湯量豊富だ, 満足だ, 無難だ, 現代的だ, 白一面だ, ···

ネガティブ語として収集された語彙の例
うっとうしい, ぎゅうぎゅうだ, すさまじい, だだっ広い, どう猛だ, ど派手だ, つまんない, はかない, むし暑い, わかりづらい, カビっぽい, ガラガラだ, システム的だ, ショックだ, ニヒルだ, 陰気だ, 壊滅的だ, 古くさい, 残虐だ, 煩い, 不親切だ, 不都合だ, 人工的だ, 五月蝿い, 単調だ, 平凡だ, 恐ろしい, 悪そうだ, 甘ったるい, 生々しい, 登りにくい, 繊細そうだ, 肌寒い, 蒸し暑い, 退屈だ, 鈍い, 零細だ, 靴下みたいだ, 高額だ, 面倒くさい, 閉鎖的だ, ···

[12] http://googlejapan.blogspot.com/2007/11/n-gram.html
[13] 奈良先端大学大学院で開発された日本語係り受け解析器 http://chasen.org/ taku/software/cabocha/
[14] 中立の語は使わない
[15] 頻度 20 以上の N-gram がエントリされている．

5.5 人手評価

極性評価辞書に無かった抽出結果について人手評価を行った．ポジティブ語，ネガティブ語として得られた語彙からそれぞれランダムに 50 語ずつ選び，合計 100 語を混ぜたデータに対して，人手で「ポジティブ」「ネガティブ」「中立」「不明[16]」の 4 種類のラベルを 5 人の被験者に付与してもらった．選択基準として「各単語が何らかの文中で別の単語に対して形容詞的に使われた場合に，どのような極性が付加されるか」という観点でラベルを選択してもらった．なお被験者は，20 歳代〜50 歳代の 5 名（男性 3 名，女性 2 名）である．

5 人の被験者の結果は「ポジティブ」を 1 点,「ネガティブ」を-1 点,「中立」と「不明」を 0 点として，各語に対する 5 人の評価結果を足し合わせる．そして，合計が 0 より大であれば「ポジティブ」，0 未満であれば「ネガティブ」，0 であれば「中立」または「不明」として，人手評価の正解データを作成した．なお，(1) 人手評価に利用した 100 語，(2) 前節の極性判定結果，(3) 人手評価による判定結果について，一覧にしたものを付録に掲載した．

この人手評価による正解データに基づき，提案手法である極性判定を評価した結果を表 7 にまとめた．ポジティブ語とネガティブ語それぞれ 50 語に対し，ポジティブ語の精度は 72%，ネガティブ語の精度は 68% であった．また，語としてノイズと思われる「不明」が 1 度でも付与された語を除いて，評価した結果，ポジティブ語の精度は 86%，ネガティブ語の精度は 85% であった．

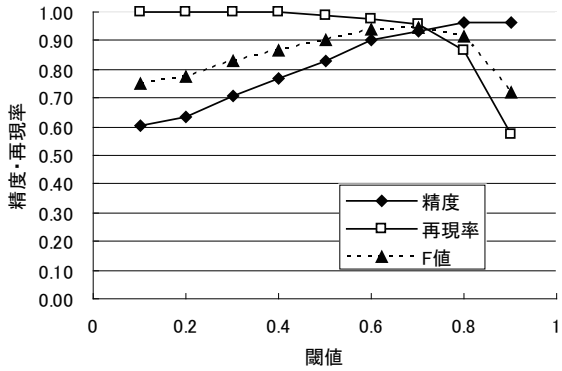

図 1: 候補語 X のうち，極性辞書にポジティブ, ネガティブ, 中立のいずれかに存在した語に関して，ポジティブのみを正解とした場合の閾値ごとの精度・再現率・F 値.

表 6: ポジティブ語・ネガティブ語収集結果

	総数	未登録語
ポジティブ語	1286 語	921 語
ネガティブ語	818 語	565 語

表 7: 人手評価による精度測定（() 内は語数の割合.）

	P の精度	N の精度
100 語の集計	0.72(36/50)	0.68(34/50)
不明が 1 度も付かない語のみ	0.86(36/42)	0.85(34/40)

5.6 考察

自動判定した極性データについて人手評価した結果，精度 72% であった．精度良く集められていると同時に，多様な細かい表現が収集できたことがポジティブ語として収集された語彙で例示した語からも分かる．

一方，間違って判定された事例を分析すると主に次の 3 種類があった．(1)「不明」ラベルが付与されるような解析時の単語抽出間違い，(2)「中立」ラベルが付与された極性を持たず性質をただ現すような単語，(3) 極性自体を反対にしてしまった単語である．(1) 単語抽出間違いでは，ひらがなの単語に多く Cabocha による解析時に単語の割り当てを間違ってしまう事例が多い．(2)「中立」ラベルについては，本手法では 2 つの極性のいずれかに割り振る閾値設定にしたため，本手法では判断できない．さらに，(3) 極性自体が反対になった単語については，大きく 2 つの問題があるといえる．1 つ目は，N-gram データ中の出現回数が少ない単語で正しく割り振れなかった事例 (例 :「あまったるい」「きよい」「探しにくい」) である．また，2 つ目は本手法が文脈を利用することが原因で，例えば候補語（構文パターンの前半の語）が極性語（構文パターンの後半の語）の程度を形容する語であって単語そのものだけになると極性が違ってしまうような事例 (例 :「衝撃的だ」「一種異様な」) があった．前者に対しては構文パターンを利用するため出現頻度が限られてしまう点を改善するような方法，後者に対しては文脈に依存に依存しない判定方法を組み合わせるなどの対応が必要といえる．

6 旅行情報の口コミデータからの理由収集

本章では，実際の旅スケの全国の口コミ投稿記事からの理由収集結果について説明する．まず目的データの理由情報の自動収集のため，ルールによらない理由タイプについて辞書を用意する．辞書作成時には，旅スケの全国 47 都道府県の口コミデータをダウンロードしたデータ（計 40645 記事，252692 文）を解析して前述した手法により語彙を集めた．その結果，(a-ii) 登録名として 104 語，(e) 対象者の限定として 131 語集まった．また，(d) ポジティブな評価については，既存の極性評価辞書と前章で生成したポジティブ評価語辞書を利用する．(a-i) と (b) は既に説明した判定ルールに従い判定する．

(a-ii) 登録名 (全 104 語の一部)
重要文化財, 天然記念物, 国宝, 世界遺産, 史跡, 文化財, 特別名勝, 名勝, 登録有形文化財, 重要伝統的建造物群保存地区, 有形文化財, 西海国立公園, 北海道遺産, 特別天然記念物, 登録文化財, 重要無形民俗文化財, 重文, 国宝・重要文化財, 海中公園, 百選, 千葉県有形文化財, 重要有形文化財, 重要無形文化財, 重要伝統的建造物保存地区, 需要文化財, 指定有形文化財, 国定公園, 国指定天然記念物, 歴史的建造物, 銘柄産地, 名水, 名所文化財, 北海道指定有形文化財, 保存地区, 福岡県有形文化財, 特別名称築山, 特別保護地区, 道 100 選, 登録文化財建造物, 登録文化財, ···

(e) 対象者の限定 (全 131 語の一部)．理由判定時には，文中で「〜向け」「〜向き」「〜用」「〜にお勧め」のいずれかのパターンの前にこれらの語が利用された場合にのみ，理由として判定する．
30 代, 2 次会, お子さま, お子さん, お子様, お祝い, お忍び, お母さん, こども, ちびっこ, ウチナー, オトナ, カップル, クリスマスデート, コドモ, ゴルフ場利用者, デート, ナイチャー, バックパッカー, パーク, ビギナー, ビジネス, ビジネスマン, ビジネスランチ, ファミリー, ファミリー・初心者, プロ, ホテル滞在者, ボード初心者, マニア, ママ, リピーター, レディース, 一般, 飲み会, 宴会, 家族, 家族連れ, 会員, 外国人, 外国人観光客, 外人, ···

そして，上記旅スケデータから理由抽出を行った．その結果，98776 の理由が抽出され[17]，全体の約半数の文から理由が獲得された．なお，3.3 節で清水寺の口コミ文から目的データを人手で抽出した際には，対象文のうちのおよそ $\frac{1}{3}$ の文から目的データが取得されていた．目的データ生成には，理由の他に対象や行動の項目も取得する必要があるが，必須項目の中心的要素である理由が全国の口コミ文の半数から得られた点で

[16] 単語として意味が分からないノイズデータ
[17] 1 文から複数の理由が得られることもある

は，目的データ生成に必要な箇所の検出を十分できたと考えられる．

理由タイプごとに抽出数と抽出例を表 4 にまとめた．理由タイプのうち，(a) 特異性では (a-i)『数詞を伴う表現』と (a-ii)『登録名』で，それぞれ 174 事例と 2695 事例あった．(b) 場所から連想されるイメージとしては，359 事例あった．(d) ポジティブな評価としては，極性辞書を元に判定した 63931 事例のほか，極性辞書にない語をポジティブ評価語判定で語彙拡張した辞書で判定した 30804 事例あった．(e) 対象者の限定としては 813 事例あった．

全体として，圧倒的にポジティブな表現による理由が多いことが分かる．また既存の極性評価辞書は，一般的によく出る表現をカバーしており獲得した理由のうち 6 割を占めた．一方，提案したポジティブ評価語の語彙獲得手法により作成した辞書により，極性辞書に含まれないポジティブな語として理由全体の 3 割を獲得できたことが分かり，辞書拡張の効果が確認できた．また，残りの 1 割はポジティブ評価以外の理由であるが，具体的な情報に基づく理由でありユーザとって有用な情報となると考えられる．

さらに，清水寺の口コミ情報のうち目的データが人手抽出できた 99 文を対象に，自動で理由抽出を行った．その結果，人手抽出した理由数 120 箇所に対し，93 箇所となった．また，表 5 に清水寺の口コミ情報から自動で抽出した理由タイプ別分類結果を示した．人手で分類した表 2 の結果と比較すると，(b) 場所から連想されるイメージや (c) 特徴的は (d) ポジティブな評価の辞書で拾われることが多く，割合として (d) が大多数を占める．一方 (a) や (e) のような表現は，ポジティブ評価語としては集めにくいが，それぞれの辞書で半数程度集めることができていた．なお本結果から自動抽出結果のうち，既存の極性評価辞書で 74% の理由箇所を獲得され，提案手法のポジティブ評価語獲得手法により 17% をさらに集めることができた．

7 おわりに

本稿では，旅行情報を例題としてデータの分析，目的データの定義を行うとともに，目的データに必要な要素として「理由」に注目し分類し，各タイプごとの理由抽出方法を提案した．さらに，「理由」の分類のうち最も表現のバリエーションが多かった「ポジティブな表現」をより多く抽出できるように，対象文書とした口コミデータ中に現れる表現を極性判定し，既存の辞書である極性評価辞書を拡張する方法を提案した．本手法では，事前に口コミ情報から形容詞的な表現を候補語として用意し，候補語と極性評価辞書中の語からなる特定の構文パターンを満たした Google N-gram のフレーズの頻度情報を利用することによって，候補語がポジティブかを判定している．その結果，新たに得られたポジティブな評価語彙は 921 語あり，人手による評価の結果では精度 70% を確認した．さらに，実際の旅行情報の口コミデータ約 4 万記事から本稿で定義した理由分類ごとに理由抽出を行い，約 10 万事例の理由が得られた．この結果，本口コミから極性評価辞書を使った理由抽出が約 6 万事例に対し，新たに得られた語彙による理由抽出が約 3 万事例あり幅広い表現に対応できたことが分かる．今後は，得られた目的データを使って，検索者の意図する目的に合わせた検索ができるよう目的データ同士の関係性の抽出を行う予定である．

参考文献

[1] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[2] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proc. of ACL*, pp. 174–181, 1997.

[3] Jian Hu, Gang Wang, Fred Lochovsky, Jian T. Sun, and Zheng Chen. Understanding user's query intent with wikipedia. In *Proc. of WWW*, pp. 471–480, 2009.

[4] Kentaro Inui, Shuya Abe, Hiraku Morita, Megumi Eguchi, Asuka Sumida, Chitose Sao, Kazuo Hara, Koji Murakami, and Suguru Matsuyoshi. Experience mining: Building a large-scale database of personal experiences and opinions from web documents. In *Proc. of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 314–321, 2008.

[5] Jaap Kamps, Robert J. Mokken, Maarten Marx, and Maarten de Rijke. Using wordnet to measure semantic orientation of adjectives. In *Proc. of LREC 2004*, pp. 1115–1118, 2004.

[6] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Expanding domain sentiment lexicon through double propagation. In *Proc. of IJCAI-09*, pp. 1199–1204, 2009.

[7] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, Vol. 37, No. 1, 2011.

[8] Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Building bridges for web query classification. In *Proc. of SIGIR*, pp. 131–138, 2006.

[9] Peter Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. of ACL*, pp. 417–424, 2002.

[10] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, pp. 201–241, 2006.

[11] 工藤拓, 賀沢秀人. Web 日本語 N グラム第 1 版. 言語資源協会発行.

[12] 高野敦子, 池奥渉太, 北村泰彦. 因果関係に着目した口コミ web サイトからの評価表現抽出. 人工知能学会, Vol. 24, No. 3, pp. 322–332, 2009.

[13] 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, 第 12 巻.

[14] 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, 2008.

[15] 倉島健, 藤村考, 奥田英範. 大規模テキストからの経験マイニング. 電子情報通信学会論文誌 D, Vol. J92-D, No. 3, pp. 301–310, 2008.

[16] 東山昌彦, 乾健太郎, 松本裕治. 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会第 14 回年次大会論文集, pp. 584–587, 2008.

[17] 那須川哲哉, 金山博. 文脈一貫性を利用した極性付評価表現の語彙獲得. 情報処理学会研究報告 2004-NL-168, pp. 109–116, 2004.

[18] 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. 情報処理学会自然言語処理研究会 (NL-144-11), 2001.

付録 ～極性判定された単語の人手評価～

次の表に人手評価を行った 100 語の単語[18]と，提案手法による極性判定の結果と，人手評価による結果の一覧を載せた．人手評価対象の 100 語，極性判定の結果，人手評価結果の一覧．（アンケート時はランダムに提示したが，ここでは見易さのため人手評価結果でポジティブらしい順に並べた．） 左から順に，単語（ポジティブ語やネガティブ語で収集された語からランダムに 50 語ずつ選んだ単語），極性判定結果としてはポ

[18] 終止形で示した．

ジティブ (P)・ネガティブ (N) の判定した結果，人手評価の結果は 5 人の結果を元に生成した正解，人手評価値は実際のラベルの合計値 (P は 1 点，N は-1 点，その他 (E) を 0 点で 5 人の総和)，不明判定数は不明ラベルを付与した人の数を表す．

語彙	極性判定結果	人手評価結果	人手評価値	不明判定数
ここちよい	P	P	5	
にこやかだ	P	P	5	
バラエティー豊かだ	P	P	5	
ポピュラーだ	P	P	5	
絢爛豪華だ	P	P	5	
隠れ家的だ	P	P	5	
可愛いらしい	P	P	5	
観やすい	P	P	5	
機能的だ	P	P	5	
気持ちよさそうだ	P	P	5	
芸術的だ	P	P	5	
小綺麗だ	P	P	5	
親切丁寧だ	P	P	5	
選びやすい	P	P	5	
平和そうだ	P	P	5	
優しげだ	P	P	5	
平和だ	P	P	4	
きよい	N	P	4	1
ソフトだ	P	P	4	
活発的だ	P	P	4	1
幸いだ	P	P	4	
正しい	P	P	4	
雪国らしい	P	P	4	
一途だ	P	P	3	
広いようだ	P	P	3	
多感だ	P	P	3	
天然っぽい	P	P	3	
動物好きだ	P	P	3	
実戦的だ	P	P	2	
手ぶらだ	P	P	2	
衝撃的だ	N	P	2	
大阪っぽい	P	P	2	
マニアっぽい	P	P	1	
メーカーらしい	P	P	1	
安そうだ	P	P	1	
決定的だ	N	P	1	
人らしい	P	P	1	2
平らだ	P	P	1	
立体的だ	P	P	1	
いるらしい	P	E	0	2
かたい	N	E	0	
からい	N	E	0	
くらいだ	N	E	0	4
ない的だ	P	E	0	5
なりがちだ	N	E	0	3
なるらしい	N	E	0	3
みたいなのだ	N	E	0	3
わからだ	P	E	0	5
一時的だ	N	E	0	
居るみたいだ	N	E	0	2
競馬好きだ	N	E	0	
見てるみたいだ	P	E	0	2
始めたらしい	P	E	0	2
斜めだ	N	E	0	
小さいのだ	N	E	0	
青い	P	E	0	
大幅だ	N	E	0	1
無いようだ	N	E	0	2
いかがだ	N	N	-1	3
すさまじい	N	N	-1	
ないようだ	N	N	-1	1
むらだ	P	N	-1	4
遠慮がちだ	N	N	-1	
重い	N	N	-1	
短い	N	N	-1	
適当だ	N	N	-1	
風変わりだ	P	N	-1	
眠い	N	N	-1	
きわどい	N	N	-2	
だだっ広い	N	N	-2	
まちまちだ	N	N	-2	
皆無だ	N	N	-2	
残り少ない	P	N	-2	
短いのだ	N	N	-2	
せつない	N	N	-3	
遠いようだ	P	N	-3	
好戦的だ	N	N	-3	
いかめしい	P	N	-4	
偉そうだ	N	N	-4	
一種異様だ	P	N	-4	
荒い	N	N	-4	
焦げ臭い	N	N	-4	
遅かったのだ	N	N	-4	
あまったるい	P	N	-5	
あわただしい	N	N	-5	
うっとうしい	N	N	-5	
つまんない	N	N	-5	
わかりづらい	N	N	-5	
陰気だ	N	N	-5	
壊滅的だ	N	N	-5	
古くさい	N	N	-5	
残虐だ	N	N	-5	
弱い	N	N	-5	
探しにくい	P	N	-5	
煩い	N	N	-5	
不親切だ	N	N	-5	
閉鎖的だ	N	N	-5	
歩きづらい	N	N	-5	
良くないようだ	N	N	-5	
埃っぽい	N	N	-5	