# 出来事の望ましさ判定を目的とした語彙知識獲得*

乾孝司　乾健太郎　松本裕治
奈良先端科学技術大学院大学 情報科学研究科
{takash-i,inui,matsu}@is.aist-nara.ac.jp

## 1　はじめに

情報検索や情報抽出に代表されるように，膨大な量のテキストデータから何らかの有益な情報を獲得する技術の確立が強く求められている．特に近年では，テキストから意見を収集したり，例文 (1) のように，意見中の出来事が望ましい（以下，*positive* あるいは $p$ ）出来事か望ましくない（以下，*negative* あるいは $n$ ．また，以下では望ましさに関する属性値を $pn$ 値と呼ぶ）出来事であるかを自動判定する技術に対する需要が高まりつつある（例えば，文献 [8, 9, 6]）．

(1) a. 美味しくて良かったよ。$\langle p \rangle$
　　b. せっかく来たのにお店が閉まっていた。$\langle n \rangle$

ここで，"良い/*positive*" や "悪い/*negative*" のような，$pn$ 値が付与された語彙知識集合（以下，$pn$ 辞書）は，出来事（を表す文）の望ましさを判定する際に有益な情報となり得る．しかしながら，望ましさは，取り扱う対象領域に強く依存し，領域毎に個別に辞書を構築する必要がある．そのため，辞書構築時にかかる人手負荷をできるだけ排除した，効率的な辞書構築手法が求められている．

我々は，人手処理を介在させずに，初期 $pn$ 辞書からブートストラップ的に $pn$ 辞書を拡張する手法を検討している．本手法は，図 1 に示すように，3 つの過程から構成され，コーパスから収集したテキスト・セグメント間における望ましさの一致関係を手がかりとして用いる．本稿では，$pn$ 辞書拡張手法の概要を説明し，提案手法の実現可能性に関する現状の調査結果を報告する．

## 2　望ましさの定義

ある出来事の望ましさは，それがどのような対象領域であるか，主体がその出来事とどのような側面や視点から関わるかに依存する．本稿では，出来事の望ましさを次のように定義する．

- 出来事の望ましさ： 主体間で共通する一つの目標状態を仮定する．もし，目標に到達する，あるいは目標到達に近付いているならば，その出来事を *positive* とし，逆に，目標の達成から遠のくような出来事のことを *negative* とする．どちらにも該当しない出来事は中立（*neutral* あるいは $e$ ）とする．

*Acquiring lexical knowledge for event desirability analysis
INUI Takashi, INUI Kentaro and MATSUMOTO Yuji
Graduate School of Information Science, Nara Institute of Science and Technology

表 1: リサイクル領域の $pn$ 辞書エントリ例

| | |
|---|---|
| *positive* | 楽しい　回収　再生品　分別<br>分別作業　ごみ減量　低公害車 |
| *negative* | 恐れる　ゴミ問題　深刻化　回収コスト<br>捨てる　過剰包装　環境破壊　無駄 |

本稿では，環境保全を目標状態と定め，リサイクル関連の新聞記事を取り扱う．この場合，ごみの削減や紙の再利用などが *positive* な出来事となる．(2)，(3) にリサイクル領域に現れる例文を示す．表 1 は，リサイクル領域の $pn$ 辞書エントリの例である．

(2) a. スチール缶の再資源化率を引き上げる。$\langle p \rangle$
　　b. 店の袋を使わない消費者を優遇する。$\langle p \rangle$

(3) a. ＯＡ化が進み紙ごみが急増する。$\langle n \rangle$
　　b. イメージダウンを恐れ簡易包装には消極的だ。$\langle n \rangle$

## 3　$pn$ 辞書

### 3.1　エントリの特徴

我々が想定する $pn$ 辞書は次のような特徴をもつ．

- 「楽しい」のように広範な領域で現れるエントリを含む一方で「回収」や「環境破壊」のように，領域に固有なエントリを含む．
- 複合語自身とそれを分解した構成要素の両方を辞書登録する．これは，"回収ペットボトル/*positive*" と "使用済みペットボトル/*negative*" のように，同一の主辞形態素をもつにも関わらず複合語としての $pn$ 値が異なるエントリが存在することによる．
- モノ概念を表す語は，以下のように，存在や増加といった出来事として拡大解釈し，$pn$ 値を特定する．
  - 『モノ』がある（例：ごみがある）
  - 『モノ』が増える（例：ごみが増える）

  例えば「ごみがある（増える）」は *negative* として解釈でき，"ごみ/*negative*" を辞書に登録する．

### 3.2　$pn$ 演算子

表 2 に示す程度や開始，継続などを表す表現はそれ単独では $pn$ 値が定まらない [7]． 本稿では，このような表現を $pn$ 演算子と呼び，$pn$ 値をもつ通常のエントリとは別に取り扱う．演算子は，スロットを持つパタンを構成し，ある要素がスロットのフィラーとして埋まることで始めて $pn$ 値が特定できる． (4a) の「高い」は，単独では $pn$ 値が特定できないが，スロットに *positive* である「リサイクル率」を伴うことで組合せとして *positive* であると特定できる．同様に， (4b) は組合せとして *negative* であると特定できる．

表 2: $pn$ 演算子の例

| | |
|---|---|
| $plus$ | 高い　増える　乗り出す　進める　取り組む |
| $minus$ | 減らす　少ない　抑制　なくす　抑える　~ない |

(4) a. $\boxed{\text{リサイクル率}_p}$ が 高い$_{plus}$ 　$\langle p \rangle$

b. $\boxed{\text{リサイクルコスト}_n}$ が 高い$_{plus}$ 　$\langle n \rangle$

c. $\boxed{\text{環境負荷}_n}$ が 少ない$_{minus}$ 　$\langle p \rangle$

d. $\boxed{\text{株価}_e}$ が 高い$_{plus}$ 　$\langle e \rangle$

（単語右下の（$p|n|e$）は単語の $pn$ 値を表す）

$pn$ 演算子は $pn$ 値ではなく $plus$ か $minus$ の 2 極の属性値をもつエントリとする．属性値が $plus$ であればフィラーと $pn$ 演算子の組合せの $pn$ 値は，フィラーの $pn$ 値を継承し，属性値が $minus$ であればフィラーの $pn$ 値を反転させた値とする（ (4c) ）. フィラーが $neutral$ である場合は，演算子の属性値にかかわらず組合せの $pn$ 値は $neutral$ とする（ (4d) ）. 後述するパタンの文脈制約を満たさない場合，$pn$ 演算子は $neutral$ の $pn$ 値をもつ通常のエントリとして扱う．式 (1) に本稿で用いたパタンを示す．スロットは $pn$ 演算子の係り元文節にあり，助詞を介して演算子表現に係るものとする．

$$\boxed{\text{スロット}}\ (\text{が}\,|\,\text{を}\,|\,\text{に}\,|\,\text{の}\,|\,\text{は}) \rightarrow\ pn\ \text{演算子} \qquad (1)$$

（“ → ” は係り受け関係を示す）

本稿では，表 1 や表 2 のような語彙知識集合のことを $pn$ 辞書と呼ぶが，$pn$ 演算子は一般的な語クラスから成り，領域間で共通していると想定できる．そこで本稿では，$pn$ 演算子は獲得対象からは除外し，$pn$ 値を属性値としてもつ通常のエントリの獲得のみを目指す．

## 4　提案手法

本節では，人手処理を介在させずに，初期 $pn$ 辞書からブートストラップ的に $pn$ 辞書を拡張する手法について述べる．本手法は図 1 のように 3 つの過程からなる．

まず，図中の「リサイクル運動を推し進める」のように，あるテキスト・セグメント（以下，節）の構成内容語が初期 $pn$ 辞書に登録されている時，構成語の組合せから，節の $pn$ 値は $positive$ であると推測できる（図 1 の（1）（2））. このような節を $pn$ 既定節と呼ぶ．次に，$pn$ 既定節の $pn$ 値が，$pn$ 値の定まっていない $pn$ 未定節（例えば「環境保全への関心が高まる」）と一致することがわかれば，$pn$ 既定節を経由して $pn$ 未定節の $pn$ 値が特定できる（図 1 の（3））（3）によって新たに得られた $pn$ 既定節の構成内容語の中に，現在の $pn$ 辞書エントリに存在しないものがあれば（例えば「環境保全」），$pn$ 辞書と節の $pn$ 値を利用して，単語の $pn$ 値を判定し，新エントリとして $pn$ 辞書に登録する（図 1 の（4）と（5））. 新規登録により更新された $pn$ 辞書を用いて，新たな $pn$ 未定節の $pn$ 値判定を試み，新規登録単語が無くなるまで（1）~（5）の処理を繰り返す．

$pn$ 構成，$pn$ 遷移，$pn$ 分解の個々の過程の実現可能性について調査した．以下，各々について報告する．

### 4.1　$pn$ 遷移過程

#### 4.1.1　$pn$ 一致の同定

$pn$ 遷移過程では，節間における $pn$ 値の一致関係を利用し，$pn$ 既定節から $pn$ 未定節へ $pn$ 値を遷移させる．この過程を実現するには，$pn$ 値の一致関係が成立する節の対を同定する手法が必要となる．

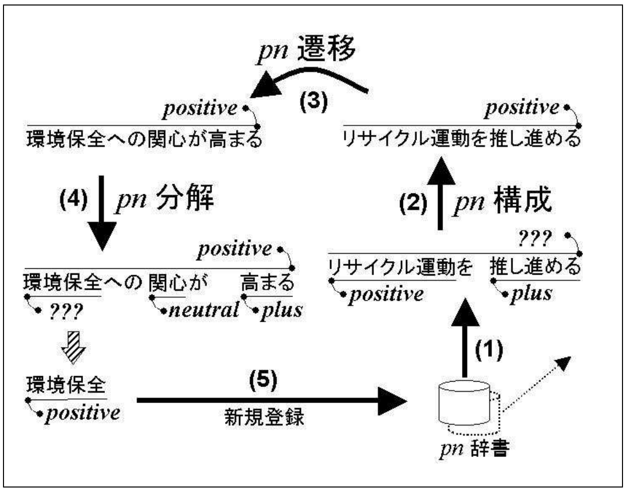


図 1: 処理のながれ

小林ら [4] の国語辞典を利用した $pn$ 辞書構築手法では，辞典中の見出し語と語釈文の双方の $pn$ 値をブートストラップ的に求め，$pn$ 辞書を構築する．この手法では「見出し語と語釈文の間では $pn$ 値が一致する」という仮定に基づき $pn$ 一致関係を得ているが，この場合，被覆率が確保できないという問題があった．

我々は，知識源としてコーパスを利用し，コーパスから抽出した因果関係に立つ出来事対から $pn$ 一致関係を得ることを考える．これは因果関係に立つ出来事間の $pn$ 値は一致するという次の仮説に基づく．

- 仮説：$positive$ な出来事は，$positive$ な出来事を引き起こす．また，$negative$ な出来事は，$negative$ な出来事を引き起こす．

もし上記の仮説が正しければ，文献 [3] で示した手法により因果関係に立つ出来事対を大量に収集することによって，$pn$ 一致関係をもつ節の対が自動的かつ大規模に獲得できる可能性がある．

#### 4.1.2　検証

新聞記事 [1, 2] 中のリサイクル関連記事から獲得した因果関係に立つ出来事対 539 件を対象にして，2 つの出来事間での $pn$ 値の一致率を調査した．表 3 に新聞記事に含まれていた因果関係に立つ出来事対の例，関係毎の出現頻度と $pn$ 値が一致していた事例数をそれぞれ示す．因果関係は文献 [3] に従って 4 種に分類した[1]．

関係種類により一致率は異なるが，$cause$ 関係，$effect$ 関係，$means$ 関係では一致傾向があった．これらの関係からは，ある程度高い割合で $pn$ 一致関係が同定できることが伺える．一方，$precond$ 関係では一致と不一致が同数程度あった．特に，表 3 の例のように「望ましくない出来事を解消するための行為」という形式での不一致が多く観測された．現状の調査は小規模ではあるが，今後必要に応じて因果関係の種類をさらに分類，

[1] それぞれの因果関係は，おおよそ次の意味関係をあらわす．“$cause$”（原因-結果），“$effect$”（効果），“$precond$”（前提条件），“$menas$”（手段）．詳細は文献 [3] に譲る．

表 3: リサイクル領域に現れる因果関係例

| 関係名 | 前件 | 後件 | 頻度 | 一致数 |
|---|---|---|---|---|
| *cause* | 天候不順により回収量が伸び悩む $\langle n \rangle$ | 段ボール古紙などの流通在庫が半分以下に落ち込む $\langle n \rangle$ | 107 | 89 |
| *effect* | 自治体がゴミ分別回収に積極的に取り組む $\langle p \rangle$ | 使用済み缶の集荷率が高まる $\langle p \rangle$ | 37 | 30 |
| *precond* | 大量の生ごみが発生する $\langle n \rangle$ | たい肥に変えてリサイクルを推進する $\langle p \rangle$ | 70 | 28 |
| *means* | 再生紙利用のガイドラインを作成する $\langle p \rangle$ | 古紙のリサイクルを促進する $\langle p \rangle$ | 325 | 316 |

$$pn_{\_tree}(i,j) = \begin{cases} pn_{\_bp}(j) & \text{if } pn_{\_bp}(j) = \; p \mid n \\ pn_{\_TREE}(i) & \text{otherwise} \end{cases}$$

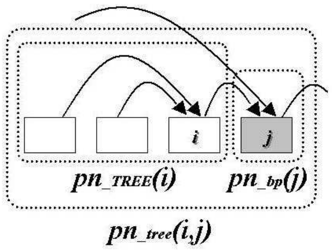


図 2: $pn$ 構成規則

foreach $i$ ( 節頭文節 ... 節末文節)
① 文節 $i$ を根とする最大係り受け木の $pn$ 値 $pn_{\_TREE}(i)$ を特定
② if $i =$ 節末文節 then 終了
③ 文節 $i$ の係り先文節 $j$ を取得
④ $pn$ 辞書を用いて文節 $j$ の $pn$ 値 $pn_{\_bp}(j)$ を特定
⑤ $pn$ 構成表を用いて $pn_{\_tree}(i,j)$ を特定，値を記憶領域に格納
end

図 3: $pn$ 構成アルゴリズム

選択的に利用することで，$pn$ 一致関係をもつ節の対がコーパスから得られると期待できる．

### 4.2 $pn$ 構成過程

#### 4.2.1 手法

$pn$ 構成過程では，$pn$ 辞書を利用して節の $pn$ 値を判定する．判定アルゴリズムは，部分係り受け木の $pn$ 値を推定する規則を用いて，係り受け木の $pn$ 値を段階的に構成することによって節の $pn$ 値を求める．

$pn$ 構成規則： 本規則は「主辞要素ほど節全体の $pn$ 値に強く影響を与える」という仮説に基づく．この仮説を係り受け関係に対して考えるならば，これは「係り側文節よりも受け側文節ほど節の $pn$ 値に強い影響を与える」と解釈できる．例えば，例文 (5) では，下線部の要素が節全体の $pn$ 値に最も強く影響すると考えられる．

(5) a. 再生$_p$が → <u>難しい</u>$_n$ → 材料$_e$を → 含む$_e$ $\langle n \rangle$

b. 深刻化$_n$する→ゴミ問題$_n$に→<u>対処する</u>$_p$ $\langle p \rangle$

（“→” は係り受け関係を示す）

構成規則を図 2 に示す．文節 $i$ は文節 $j$ に係っており，$pn_{\_bp}(j)$ は文節 $j$ の $pn$ 値，$pn_{\_TREE}(i)$ は，文節 $i$ を根とする最大係り受け木の $pn$ 値，$pn_{\_tree}(i,j)$ は文節 $i$ を根とする最大係り受け木に文節 $j$ を加えた部分木全体の $pn$ 値を表す．$pn$ 構成規則および 3.2 節で述べた $pn$ 演算子から導出される $pn_{\_tree}(i,j)$ は，表 4 のように $pn$ 構成表としてまとめられる．

表 4: $pn_{\_tree}(i,j)$ を求める $pn$ 構成表

| | | $pn_{\_bp}(j)$ | | | | |
|---|---|---|---|---|---|---|
| | | $p$ | $n$ | $plus$ | $minus$ | $e$ |
| $pn_{\_TREE}(i)$ | $p$ | $p$ | $n$ | $p$ | $n$ | $p$ |
| | $n$ | $p$ | $n$ | $n$ | $p$ | $n$ |
| | $e$ | $p$ | $n$ | $e$ | $e$ | $e$ |

表 5: $pn$ 構成による節の $pn$ 値判定精度

| | 再現率 | | 適合率 | |
|---|---|---|---|---|
| $p$ | 0.90 | (656/725) | 0.99 | (656/665) |
| $n$ | 0.87 | (104/119) | 0.79 | (104/132) |
| $e$ | 0.60 | ( 12/ 20) | 0.18 | ( 12/ 67) |

$pn$ 構成アルゴリズム ( 図 **3** ): $pn$ 構成表に従って，節頭文節から順に係り受け木の $pn$ 値を順次判定していき，最終的には，$pn_{\_TREE}$(節末文節) を節の $pn$ 値とする．文節の $pn$ 値は，格助詞などの補足語は考慮せず，内容語の $pn$ 値で代表させ，内容語の $pn$ 値は $pn$ 辞書を引くことで特定した．図 3 の操作⑤では，受け側文節 $y$ とそのすべての係り側文節 $x$ から得られる $pn_{\_tree}(x,y)$ を，求めた順にリスト $list(y)$ に格納する．リストは文節毎に用意し，文節 $y$ のリスト $list(y)$ は，操作① で $pn_{\_TREE}(y)$ を特定する際に利用する．具体的には，$list(y)$ の最も末尾側にある $positive$ か $negative$ を $pn_{\_TREE}(y)$ とした．

#### 4.2.2 評価実験

提案手法による節の $pn$ 値判定精度を実験的に検証した．実験の設定を以下に示す．

- 対象データは 4.1.2 節で用いた因果関係の出来事対 539 件のうち，$cause$ 関係，$means$ 関係の 432 件，864 節 ( 平均 5.2 文節 / 節 )．
- 864 事例に含まれるすべての単語を洩れなく収録した $pn$ 辞書を人手で構築して使用した．エントリ数はそれぞれ，$positive/1126$，$negative/489$，$plus/124$，$minus/52$，それ以外は $neutral$ である．
- 「～ない」，「～にくい」は独立した文節として扱う．

表 5 に判定精度を示す．提案手法は，係り受け関係に基づく単純な手法ではある[2] が，新聞記事中から得たリサイクル領域の節に関する限り，$positive$ と $negative$ に関しては高い精度を得ており，本手法の有効性を示唆している．本アルゴリズムでは対処できない典型的な誤り事例を例と共に以下に示す．まず，$pn$ 値が $positive$ か $negative$ である単語がひとつでも節内に含まれていれば，節の $pn$ 値を $neutral$ と判定できない（ (6a) ）．逆に，節内に含まれている単語の $pn$ 値がすべて $neutral$ であれば，節の $pn$ 値を $positive$ あるいは $negative$ と判定できない（ (6b) ）．また，省略表現（ (6c) の「値段の」）や連体修飾（ (6d) の「再使用可能な」）を含む

[2] 例えば，目良 [5] の情緒値計算式は，格役割を考慮するなど，我々の手法を精緻化したものと捉えることができる．

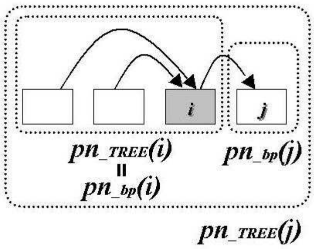


図 4: $pn$ 分解

節での誤りが比較的多く観測された．今後は対象領域，データ規模を増やしつつ，これらの誤りに対応したアルゴリズムの開発を進めていく予定である．

(6) a. 【$neutral$ と判定できない場合の誤り例】
排水$_n$の 再利用$_p$を 繰り返す$_e$と 水分$_e$の イオン濃度$_e$が 上がる$_{plus}$
正答=$\langle neutral\rangle$ 出力=$\langle positive\rangle$

b. 【$neutral$ と判定した場合の誤り例】
微生物$_e$の 働き$_e$で 土$_e$に 戻る$_e$
正答=$\langle positive\rangle$ 出力=$\langle neutral\rangle$

c. 【省略表現を含む誤り例】
再生紙$_p$の( 値段$_n$の )方$_e$が 上質紙$_e$より 高い$_{plus}$
正答=$\langle negative\rangle$ 出力=$\langle positive\rangle$

d. 【連体修飾を含む誤り例】
再使用可能$_p$な 包装紙$_n$の 普及$_{plus}$を 目指す$_{plus}$
正答=$\langle positive\rangle$ 出力=$\langle negative\rangle$

### 4.3 $pn$ 分解過程

#### 4.3.1 手法

$pn$ 分解過程では，$pn$ 辞書と節の $pn$ 値を利用し，節内の単語の $pn$ 値を特定する．具体的には，表 4 に代わり，表 6 の $pn$ 分解表を用い，構成過程の逆の操作を行う．ある受け側文節 $j$ に対して最も節末に位置する係り側文節が $i$ である時，$pn\_{TREE}(j)$ と $pn\_{bp}(j)$ の値を利用して文節 $i$（の内容語）の $pn$ 値 $pn\_{bp}(i)$ を判定する（図 4）．$pn$ 分解表によって得られる $pn$ 値は $pn\_{TREE}(i)$ であるが，受け側文節優先仮説を踏襲し，ここでは $pn\_{TREE}(i)$ で $pn\_{bp}(i)$ を近似する．$pn$ 分解表に従って，節末から節頭に向かって処理を進め，$pn\_{TREE}(i)$ が特定できなった時点で処理を終了する．

#### 4.3.2 評価実験

4.2.2 節と同一のデータ，条件設定で $pn$ 分解過程の精度を検証した．表 7 左側（“制約なし”）に判定精度を示す．また，判定できたエントリ例を以下に示す．

- 分別回収$_p$ 食い止める$_p$ 再利用率$_p$ 循環型社会$_p$
- 回収漏れ$_n$ 伸び悩む$_n$ 生ごみ廃棄物$_n$ 過剰包装$_n$

表 7 から，$pn$ 値が $neutral$ である単語に対して，誤った $pn$ 値を割り当てるケースが多いことがわかる．そこで，次の制約を新たに課して分解過程を実行した．

- 分解制約：分解過程で新たに $pn$ 値が特定された単語について，$pn$ 値を $e$ に置き換えて構成過程を実行する．この時，もし節の $pn$ 値が正しく特定されるならば，分解過程の出力を保留する．

表 6: $pn\_{TREE}(i)$ を求める $pn$ 分解表

| | | $pn\_{bp}(j)$ | | | | |
|---|---|---|---|---|---|---|
| | | $p$ | $n$ | $plus$ | $minus$ | $e$ |
| $pn\_{TREE}(j)$ | $p$ | − | − | $p$ | $n$ | $p$ |
| | $n$ | − | − | $n$ | $p$ | $n$ |
| | $e$ | − | − | $e$ | $e$ | $e$ |

表 7: $pn$ 分解による単語の $pn$ 値判定精度

| | 制約なし | | 制約あり | |
|---|---|---|---|---|
| $p$ | 0.92 | (157/170) | 0.94 | (124/132) |
| $n$ | 0.71 | ( 44/ 62) | 0.74 | ( 32/ 43) |
| $e$ | 0.10 | ( 9/ 90) | 0.17 | ( 4/ 24) |

表 7 右側（“制約あり”）に判定精度を示す．制約下では，被覆率が低下する一方で，精度の改善が確認できる．制約によって保留された事例を (7) に示す．「家庭ごみ」が出力を保留された対象語である．

(7) 回収$_p$を 家庭ごみ に 広げる$_e$ $\langle p\rangle$
正答=$\langle negative\rangle$ 制約なしでの出力=$\langle positive\rangle$

上記の分解過程の基本原理は構成過程に依る部分が大きい．そのため，制約の有無に関わらず，構成過程で誤っていた事例については，同様に分解過程でも誤る傾向があった．定性的な誤り分析は今後の課題である．

## 5 おわりに

本稿では，初期 $pn$ 辞書からブートストラップ的に $pn$ 辞書を拡張する手法を提案した（図 1）．現在までの調査結果から，$pn$ 遷移過程では，因果関係に立つ出来事間で $pn$ 一致関係が成立する傾向があることを確認した．また，$pn$ 構成・分解過程では「主辞要素ほど節全体の $pn$ 値に強く影響を与える」という仮説に基づく規則およびアルゴリズムを提案し，ある程度高い精度で $pn$ 値が判定できることを確認した．ただし，ブートストラップによって自動的に $pn$ 辞書を拡張するには，各過程の精度は十分に高いとは言い難く，今後，各々の洗練を進めていきたい．

### 参考文献


[1] 日本経済新聞社. 日経産業新聞 CD-ROM 版.

[2] 日本経済新聞社. 日本経済新聞 CD-ROM 版.

[3] 乾孝司, 乾健太郎, 松本裕治. テキストから獲得可能な因果関係知識の類別およびその自動獲得の試み-接続助詞「ため」を含む文を中心に-. 言語処理学会第 9 回年次大会, pp. 707–710, 2003.

[4] 小林のぞみ, 乾孝司, 乾健太郎. 語釈文を利用した「p/n 辞書」の作成. 人工知能学会言語・音声理解と対話処理研究会 (SLUD-33), 2001.

[5] 目良和也. 語の好感度に基づく自然言語発話からの情緒生起手法. 信学技報 (NLC98-26), pp. 1–8, 1998.

[6] 村野誠治, 佐藤理史. 文型パターンを用いた主観的評価文の自動抽出. 言語処理学会第 9 回年次大会, pp. 67–70, 2003.

[7] 長江朋, 望月源, 白井清昭, 島津明. 製品コンセプトと製品評価文章の関係の分析. 言語処理学会第 8 回年次大会, pp. 583–586, 2002.

[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. 情報処理学会自然言語処理研究会 (2001-NL-144), pp. 75–82, 2002.

[9] P.D. Turney. thumbs up? thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. of the 40th. ACL*, pp. 417–424, 2002.