

意見マイニングを志向したQAサイト投稿テキストの解析

井上 結衣[†] 藤井 敦^{††}

[†] 筑波大学大学院図書館情報メディア研究科 〒305-8550 茨城県つくば市春日 1-2

^{††} 東京工業大学大学院情報理工学研究科 〒152-8552 東京都目黒区大岡山 2-12-1

あらまし World Wide Web 上には、意見、評判、感想などの主観情報が多く含まれる。複数の人間が書いた主観情報から人々の考え方に関する傾向や法則を発見することができれば、個人や組織の意思決定において有益な情報となる可能性がある。本研究は、意見マイニングにおける「意見収集」と「極性反転文検出」に関する手法を提案する。意見収集は、QA サイトから時事問題に対する意見を収集し、賛否に基づいて分類する。極性反転文検出は、意見段落中から意見全体の極性とは異なる極性で述べられている文を検出する。評価実験によって提案手法の有効性を示す。
キーワード 意見マイニング, QA サイト, 極性判定

Analyzing Articles in QA Sites for Opinion Mining

Yui INOUE[†] and Atsushi FUJII^{††}

[†] Graduate School of Library, Information and Media Studies, University of Tsukuba
Kasuga 1-2, Tsukuba-shi, Ibaraki, 305-8550, Japan

^{††} Graduate School of Information Science and Engineering, Tokyo Institute of Technology
Oookayama 2-12-1, Meguro-ku, Tokyo, 152-8552 Japan

Abstract On the World Wide Web, the volume of subjective information, such as opinions and reviews, has been increasing rapidly. The trends and rules latent in a large set of subjective descriptions can potentially be useful for decision-making purposes. In this paper, aiming at opinion mining, we propose methods for collecting opinions and detecting polarity inversion. For collecting opinions, we search question-answering sites for opinion texts and classify them based on the polarity. For detecting the polarity inversion, we identify sentences whose polarity is opposite from the polarity for the entire opinion text. We show the effectiveness of our methods experimentally.

Key words Opinion Mining, Question-Answering Site, Polarity Inversion

1. はじめに

World Wide Web には、報道記事のように客観性が高い情報だけでなく、意見、評判、感想などの主観情報も存在する。例えば、「赤ちゃんポスト」などの時事問題に対して、様々な人々が掲示板やブログ等を通して自分の考えを発信している。これらの意見情報から人々の考え方に対する傾向や法則を発見することができれば、個人や組織の意思決定に役立つ可能性がある。

筆者らは、Web から時事問題に対する意見情報をマイニングし、その傾向を可視化することによってユーザの意思決定を支援するシステム「OpinionReader (オピニオンリーダー)」[3], [6] について研究している。

本研究において、意思決定とは、ある話題に対する賛否両論を網羅的に洗い出し、対立させて、より合理的な立場を採用する過程と捉える。OpinionReader は、ある話題について賛成派

と反対派が対立する様子を、賛成または反対の根拠となる「論点」に基づいて2次元グラフ上に可視化する。

図1は、「赤ちゃんポスト」に対する出力の例である。「虐待」などの論点を2次元グラフ上に表示する。グラフの縦軸は論点の重要度を表し、横軸は論点がどれだけ賛成もしくは反対に固有かを表す。論点を選択すると、該当する論点を含む意見が順位つきリストで表示される。以上の機能により、ユーザは大量の意見情報を読まなくてもその話題に関する議論の全容を把握することができる。

システムの入力には、「赤ちゃんポスト」などの時事問題である。時事問題が与えられると、システムは以下の処理を行う。

- (1) 意見収集: Web から時事問題に対する賛成意見と反対意見を区別して収集する。
- (2) 論点抽出: 収集した意見情報から論点を抽出する。

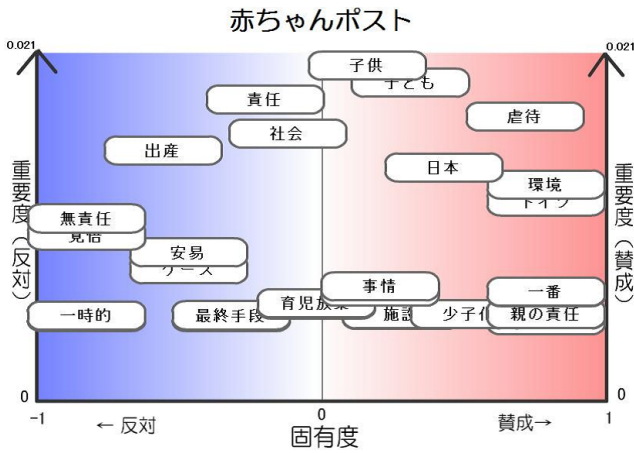


図 1 OpinionReader の出力例

(3) 可視化：固有度と重要度に基づいて論点を可視化する．

本研究は，(1)「意見収集」と(2)「論点抽出」の改善を目的とする．

上記(1)の「意見収集」では，ある時事問題に対して賛成または反対の意見を Web から自動的に収集する[3]．しかし，現在の手法では収集できる意見が少なく，また賛成と反対に分類する精度が低いという問題があった．

そこで本研究は，QA サイトから意見情報を収集する手法を提案する．QA サイトとは，ユーザが投稿した質問に対して，別のユーザが回答を投稿する形式で知識の共有を行う Web サイトである．具体例として，Yahoo!知恵袋^{注1)}や OKWave^{注2)}がある．一般的に，1つのページには1件の質問とそれに対する複数の回答が表示される．ユーザは質問や回答を投稿するだけでなく，投稿された質問と回答を検索することができる．

本手法は，QA サイトから「赤ちゃんポストに賛成？反対？」のような質問が投稿されたページを検索し，その質問に対して投稿された回答群から意見情報の収集を行う．このような質問に対して投稿される回答は，「反対です。子捨てを容認することになります。」のように，立場の表明とその根拠が書かれている場合が多いため，高い精度で賛成と反対を分類できると考えた．

上記(2)の「論点抽出」における課題は，実際には論点でない語句が論点として抽出される点である．現在の OpinionReader では，意見情報から名詞句と動詞句を論点として抽出する．しかし，実際の意見情報には，全体としては賛成でも反対の意見について言及した上で反論したり，反対の主張に譲歩している文などが含まれる．以下の例1と例2は，それぞれ「赤ちゃんポスト」に対する賛成意見と反対意見である．

例1) 僕も賛成です。これで虐待死が減るのであれば絶対にあった方がいいです。ただ、匿名では安易な捨て子に繋がると思います。

例2) 反対。捨てられて死ぬよりはマシというのわかる。でもなんか納得がいけない。違う助け方をもっと充実させればい

いんじゃないだろうかと思う。

例1)と例2)における下線部は，冒頭で表明されている投稿者の立場とは逆の立場に関する記述である．

本研究では，このような段落中で立場が反転している文を「極性反転文」と呼び，極性反転文を検出するための手法を提案する．

2. 関連研究

Web からの意見収集に関する研究では，日記やブログのように主観情報を多く含む文書を選択的に収集する手法[2]や，文書中の主観的な記述を収集する手法[1]がある．しかし，意見収集の多くはレビューなどの評価テキストを対象としているため，時事問題に対する意見を収集する研究は少ない．また，QA サイトの意見分析に関する研究として，関ら[7]の研究がある．関らは QA サイトの文書を対象として，時事問題などのトピックに対する意見情報のタイプを付与し，出現する意見情報の傾向について考察した．しかし，自動的に特定することを行わない点で本研究とは異なる．

極性反転に関する研究として，那須川ら[5]や中道ら[8]の研究がある．那須川らは，映画のレビューなどの評価文書において「しかし」などの逆接表現をきっかけに肯定，否定の極性が反転することを利用し「面白い」や「退屈」等の極性表現を学習する手法を提案した．中道らは，評価文書において情緒の極性を特定する接続表現を定義するために，文中に用いられる接続表現を「情緒の保持」「情緒の反転」「情緒の共起」の3つに分類し，極性を特定することができるか実験した．しかし，どちらの研究も極性反転文を検出することを目的としておらず，本研究とは異なる．Kimら[1]は，レビューに対して文の単位で極性を判定する手法を提案した．しかし，本研究はまず意見テキスト全体の極性を判定し，その極性とは異なる文を検出するという点で異なる．

3. 提案する手法

3.1 概要

本研究は「意見収集」と「極性反転文検出」のそれぞれに関する手法を提案する．2つの手法は有機的に連携している．まず，意見収集では時事問題に対する意見を QA サイトから抽出し，賛否に基づいて分類することで全体的な極性を判定する．次に，極性反転文検出によって部分的な逸脱を検出する．

以下の3.2と3.3で各手法について説明する．

3.2 意見テキスト収集

本手法は，QA サイトにおいて質問のタイトルに以下の2つの文字列含むページを検索する．

X 賛成(？|ですか|でしょうか)

Xは「赤ちゃんポスト」などの時事問題である．()の中は，|で区切られた記号または語句のいずれか1つに一致すれば良い．例えば「赤ちゃんポストに賛成ですか？」という文字列を質問に含むページが検索される．ただし，以下の表現を質問に含む場合は，以降の回答抽出において賛否を逆転させて扱う．

(注1)： <http://chiebukuro.yahoo.co.jp/>

(注2)： <http://okwave.jp/>

(廃止 | 撤廃 | 撤回 | 脱)

例えば、「赤ちゃんポストの廃止に賛成？」という質問に対する賛成意見は「赤ちゃんポスト」に対する反対意見である。

ページを検索したら、ページのレイアウト（HTML の構造）に基づいて回答の単位に分割し、以下の文字列を本文中のどこかに含む回答だけを抽出する。P は「賛成」または「反対」である。

P (です | します | ね | に決まってる | 派です | に 1 票 | に 1 票 | といわざるを得ません | と言わざるを得ません | せざるを得ない | 過激派です | でございます | っす | かな | だな | だね | [スペース] | [改行] | [文末記号] | だ + [文末記号])

例えば、「賛成派です。」や「反対と言わざるを得ません。」などの記述を含む回答が意見として抽出される。[文末記号] とは、以下の記号のうちのいずれかとする。「w」と「・」は、当該記号が 2 つ以上連続している場合に文末記号と見なす。「w」は、例えば「ゆとり世代ですが何かwww」のように使用されることが多い。

(。 | . | 、 | , | ! | … | ♪ | w | ・)

ただし、以下に示す文字列のみで構成されている回答は意見として抽出しない。

上記の表明表現、[付加表現]、記号

付加表現を以下に示す。付加表現は複数含んでいても、含まなくても良い。

(絶対 | 勿論 | もちろん | 大々 | 私は)

例えば、以下のような回答は意見として抽出しない。

私は大々々賛成です (^ - ^ *)

また、「賛成」と「反対」を両方とも含む回答は抽出しない。例えば、以下のような回答は抽出しない

赤ちゃんポストの主旨に賛成。実施に反対。

以上の手法により、時事問題に対して「賛成」を表明する意見情報の集合と「反対」を表明する意見情報の集合が別々に収集される。

3.3 極性反転文の検出

3.3.1 概要

極性反転文の検出では、まず、意見情報を 3.2 で示した文末記号で文単位に区切る。

次に、規則を用いて極性反転文を検出する。ただし、ここで言う極性反転文とは、文頭から文末までの全てとは限らず、任意の断片を指す。規則は極性が反転する表現を手で分析し、作成した。極性が反転する表現を言語的特徴によって「逆接」、「引用」、「譲歩」の 3 カテゴリーに分類した。ただし、表現ごとに検出する範囲が異なる。検出する範囲は以下の A か B のいずれかである。

A 文頭から手がかり表現まで

B 文頭から文末まで

同一の文に対して複数の規則が適用できる場合は、最も後ろにある表現を手がかり表現とする。以下、3.3.2 ~ 3.3.4 で各カテゴリについて説明する。

3.3.2 逆接

「逆接」は、意見文中で否定的な態度を表す場合に使用される。この規則に使用する表現は、文中のどこに出現するかによって「文頭」、「文中」、「文末」の 3 通りに分けられる。それぞれの表現と検出の範囲は以下のとおりである。なお、表現は漢字表記と平仮名表記を両方含む。

	手がかり表現	検出範囲
文頭	ただ (接続詞)、でも、だが、けど、ですが、しかし、ただし、だけれど、けれど、たとえ-でも、たとえ-としても、けれども、もつとも、しかしながら	B
文末	が (助詞-接続助詞)、けど	A
その他	が (助詞-接続助詞)、としても、反面、からといって、にせよ、けど、でも	A

「たとえ-でも」と「たとえ-としても」は「たとえ」と「ても」または「としても」の間にある文字列を検出範囲とする。() で品詞を付記した表現は ChaSen^(注3)による形態素解析結果に基づいて品詞を判定する。

例えば、以下の例における「ただ」という表現は、上記規則の「文頭」に該当する。

赤ちゃんの命が救われるので賛成です。
ただ、ネーミングは悪いと思います。

「文頭」の検出範囲は B の「文頭から文末まで」であるため、下線部を極性反転文として検出する。

3.3.3 引用

「引用」は、意見文中で他者の意見を引用する場合に使用され、「などと言う」や「という意見」が手がかり表現となる。

この規則における手がかり表現は、以下の式で表せる。

I. 一般的な表現 + (II. 修飾語) + III. 意見特有の表現

「II. 修飾語」は無くても良い。それぞれにおける表現と検出範囲は次の通りである。

I. 一般的な表現	II. 修飾語	III. 意見特有の表現	検出範囲
と、とか、って、だと、など、と、なんて、ことを	短絡的な、よく、甘い	いう (動詞)、言う、云う、仰る、コメントする、詭弁を唱える	A
との、という		意見	

引用の表現は「言う」の類義語に関する表現と「意見」とい

(注3): <http://chasen.naist.jp/hiki/ChaSen/>

う単語に関する表現の 2 パターンがある．そのため，I. と III. は 2 パターンを定義している．II. は共通である．III. の表現は活用形を含む．例えば，以下の例文における「とか言ってる」という表現は，引用の手がかり表現である．

賛成です。
育児放棄が増えるとか言ってる人もいますが、それより命が大事。

検出範囲は A の「文頭から表現まで」であるため，上記例文の下線部を極性反転文として検出する．

3.3.4 譲 歩

譲歩に関する表現は，意見文中で他者の意見や一般的な世論に譲歩している場合に使用される．この規則に使用する表現と検出の範囲は次の通りである．

手がかり表現	検出範囲
確かに，勿論，無論	B

例えば，以下の例文における「確かに」の表現が該当する．

反対です。
確かに、赤ちゃんポストがあれば命は救われる。
しかし、その後の事も考えてください。

検出する範囲は B の「文頭から文末まで」であるため，上記例文の下線部を極性反転文として検出する．

4. 評価実験

4.1 意見情報収集の評価

3.2 で示した意見情報収集の手法を用いて，Yahoo!知恵袋から意見を収集し，収集した情報に対して人手で正解判定した．対象としたトピックは「赤ちゃんポスト」「ゆとり教育」「東京オリンピック」「郵政民営化」「裁判員制度」である．意見の収集は，Yahoo!API で Yahoo!知恵袋のドメインを対象にして検索を行った．

本手法は，QA サイトにおける意見と非意見を区別し，意見のみを抽出する「抽出」と，抽出した意見を賛否に分類する「分類」の 2 つの処理から成る．そこで，表 1 と表 2 に「抽出」と「分類」それぞれに関する評価結果を示す．表 1 より高い精度と再現率で意見を収集でき，表 2 より高い精度で意見を分類できたことが分かる．

さらに「抽出」と「分類」を組み合わせた評価結果を表 3 に示す．なお，表 3 における (a)～(c) は表 1 と共通である．表 3 では「抽出」と「分類」の誤りが混在しているため，表 1 よりも結果は悪くなる．しかし，表 3 より「収集」と「分類」を組み合わせた場合でも高い精度と再現率が得られた．

表 1 に関する抽出漏れの事例を分析したところ，全て「賛成」もしくは「反対」以外の表現で立場を表明している意見が原因であった．例えば，以下の表現は「賛成」や「反対」という語句を含まないにも拘わらず反対の立場を表明している．

赤ちゃんポストなんてやめて欲しい。

表 2 に関する分類誤りの事例を分析したところ，賛成または反対の根拠が書かれていない意見や中立意見が原因であった．

また，QA サイトにおける意見の偏りについて分析した．トピックは「赤ちゃんポスト」「ゆとり教育」「東京オリンピック」の 3 つを対象とした．QA サイトから収集した意見は，Web 上の連続していない意見と比較すると，自分が投稿する前の質問や回答に影響を受けて偏っている可能性がある．そこで，松村ら [4] のコメントの媒介影響量（以下「影響量」）を用いて分析した．算出方法は，検索された各ページの回答の並びをそのままにして算出した結果と，ランダムに並び変えて算出した結果を比較した．表 4 に「質問の影響量」と「回答の影響量」の平均を示す．

表 4 から，質問の影響量は元の並びよりも若干高い値になった．そこで，影響量の差の原因を調べたところ，全てのページで差がある訳ではなく，いくつかのページで突出して高くなっていることが分かった．今後は，質問の影響量が一定以上の場合は収集する意見の重みを下げるなどの考慮が必要である．回答の影響量は元の並びとランダムでほとんど差はなかった．すなわち，QA サイトにおける回答はそれ以前に投稿された回答に影響を受けている訳ではないことが分かった．

Web 上には，今回の分析や評価に用いた「Yahoo!知恵袋」以外にも QA サイトが存在する．そこで「OKWave」を対象として「Yahoo!知恵袋」の結果と比較した．まず，2 つのサイトにおけるデータ量を比較した結果を表 5 に示す．表 5 の件数は 2010 年 3 月 19 日現在各サイトのトップページに表示されている件数である．

表 5 より「OKWave」の質問総数は「Yahoo!知恵袋」の約 8 分の 1 であり，回答総数は約 6 分の 1 である．「OKWave」のデータ量が少ない理由として，不適切な質問や回答になっていないような回答を削除する「削除規制」が厳しい点が考えられる．

次に，2 つのサイトで「赤ちゃんポスト」に対する賛否を問う質問を含む QA ページを検索し，本手法で意見収集を行った結果を表 6 に示す．表 6 における「Yahoo!知恵袋」の精度と再現率は表 3 の結果と同一である．表 6 より，質問数と回答数は「Yahoo!知恵袋」の方が多かった．しかし「OKWave」を用いた場合でも高い精度と再現率が得られた．今後は，様々な QA サイトを対象とすることで，意見収集の網羅性を上げることができる．

4.2 極性反転文検出の評価

まず，意見全体における極性反転文の割合について，4.1 と同じ 5 つのトピックを対象として調査した．結果を表 7 に示す．表 7 より，極性反転文の割合は意見文全体の約 9.2 %であった．Kim ら [1] のように文単位で極性を特定するよりも，本手法のように意見全体の極性を特定した後に極性反転文だけを検出する方が効率的である．

極性反転文の検出手法について，Yahoo!知恵袋から人手で抽

表 1 意見の「抽出」に関する評価

	赤ちゃんポスト	ゆとり教育	東京オリンピック	郵政民営化	裁判員制度
(a) 検索されたページ数	19	8	34	39	21
(b) ページに含まれる回答数	160	64	182	239	118
(c) (b) のうち調査した回答数	160	64	100	110	118
(d) (c) のうち意見数	134	52	81	79	90
(e) 自動抽出された段落数	113	44	75	71	82
(f) (d) のうち意見数	113	42	70	63	80
(g) 自動抽出の精度	100%(113/113)	95.5%(42/44)	93.3%(70/75)	88.7%(63/71)	97.6%(80/82)
(h) 自動抽出の再現率	84.3%(113/134)	80.8%(42/52)	86.4%(70/75)	79.7%(60/76)	88.9%(75/85)

表 2 意見の「極性分類」に関する評価

	赤ちゃんポスト	ゆとり教育	東京オリンピック	郵政民営化	裁判員制度
分類の精度	96.5%(109/113)	95.2%(40/42)	90.0%(63/70)	95.2%(60/63)	93.8%(75/80)

表 3 意見収集手法に関する総合評価

	赤ちゃんポスト	ゆとり教育	東京オリンピック	郵政民営化	裁判員制度
(d) (c) のうち意見数	110	54	70	76	85
(e) 自動収集された段落数	121	46	64	71	82
(f) (d) のうち意見数	110	41	59	60	75
(g) 自動収集の精度	90.9%(110/121)	89.1%(41/46)	92.2%(59/64)	84.5%(60/71)	91.5%(75/82)
(h) 自動収集の再現率	83.3%(110/132)	75.9%(41/54)	84.3%(59/70)	78.9%(60/76)	88.2%(75/85)

表 4 コメント影響量の分析

トピック	質問		回答	
	元の並び	ランダム	元の並び	ランダム
赤ちゃんポスト	0.238	0.180	0.093	0.099
ゆとり教育	0.157	0.133	0.099	0.093
東京オリンピック	0.202	0.148	0.061	0.079

表 5 「Yahoo!知恵袋」と「OKWave」におけるデータ量の比較

	Yahoo!知恵袋	OKWave
質問総数	約 3800 万件	約 480 万件
回答総数	約 9900 万件	約 1700 万件

表 6 「Yahoo!知恵袋」と「OKWave」における意見収集の比較

	Yahoo!知恵袋	OKWave
質問数	32 件	3 件
回答数	160 件	45 件
精度	90.9%(110/121)	97.0%(32/33)
再現率	83.3%(110/132)	91.4%(32/35)

表 7 意見文全体における極性反転文の割合

トピック	割合（極性反転文の数/意見文の総数）
赤ちゃんポスト	10.8%(78/721)
ゆとり教育	10.1%(20/198)
東京オリンピック	7.7%(23/299)
郵政民営化	6.1%(22/363)
裁判員制度	10.2%(34/333)
合計	9.2%(177/1914)

出した「赤ちゃんポスト」、「ゆとり教育」、「東京オリンピック」、「郵政民営化」、「裁判員制度」のそれぞれに対する意見情報を用いて評価を行った。具体的には、提案手法で検出した反転箇所に対して人手で正解判定した。結果を表 8 に示す。表 8 では、

5 つのトピック全てにおいて 80% 以上の再現率が得られた。

表 8 極性反転検出手法の評価

トピック	精度	再現率
赤ちゃんポスト	48.9%(64/131)	82.1%(64/78)
ゆとり教育	54.8%(17/31)	85.0%(17/20)
東京オリンピック	34.5%(19/55)	82.6%(19/23)
郵政民営化	38.0%(19/50)	86.4%(19/22)
裁判員制度	37.3%(28/75)	82.4%(28/34)

他方で、いずれのトピックに対しても精度が低かった。そこで、5 トピックの合計で規則ごとの精度を分析した結果を表 9 に示す。表 9 から、「逆接」に関する規則の精度が特に低いことが分かった。

表 9 極性反転文の検出に関する規則ごとの精度

	逆接	引用	譲歩
機械抽出	321	28	44
正解	135	23	29
精度	42.1%	82.1%	65.9%
	(135/321)	(23/28)	(29/44)

「逆接」における誤り 186 件の内訳を表 10 に示す。表 10 の結果は、5 トピックの合計である。

表 10 規則 (a) の誤り内訳

全体の極性と検出箇所の極性が一致する	48 箇所
極性なしの箇所	138 箇所

表 10 の結果より、誤りの多くは「極性なしの箇所」が検出された点にあった。以下の下線部はその一例である。

今日もニュースで見ましたが、3 歳児が入れられたらしいですね。

5. おわりに

本研究は、時事問題に対する議論の様子を可視化するシステム OpinionReader の改善を目的として、QA サイトを用いた意見テキスト収集と極性反転文の検出手法を提案した。今後は、対象の時事問題を増やししながら評価を繰り返し、手法の改善を行う予定である。

謝 辞

本研究の一部は、文部科学省科研費特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」(課題番号:21013003)によって実施された。

文 献

- [1] Soo-Mim Kim and Eduard Hovy. Determining the sentiment of opinions. *Proceeding of Conference on Computational Linguistics*, pp. 1367–1373, 2004.
- [2] Tomoyuki Nanno, Toshiaki Fujiki, Ysuihiro Suzuki, and Manabu Okumura. Automatically collecting monitoring and mining japanese weblogs. *In The 13th International World Wide Web Conference*, pp. 320–321, 2004.
- [3] 井上結衣, 藤井敦. Web 世論からの意見抽出と賛否に基づく分類. 言語処理学会第 14 回年次大会発表論文集, pp. 364–367, 2008.
- [4] 松村真宏, 大澤幸生, 石塚満. テキストによるコミュニケーションにおける影響の普及モデル. 人工知能学会誌, Vol. 17, No. 3, pp. 259–267, 2002 2002.
- [5] 那須川哲哉, 金山博, 坪井祐太, 渡辺日出雄. 好不評文脈を応用した自然言語処理. 言語処理学会第 11 回年次大会講演論文集, pp. 153–156, 2005.
- [6] 藤井敦. OpinionReader: 意思決定支援を目的とした主観情報の集約・可視化システム. 電子情報通信学会論文誌, Vol. J91-D, No. 2, pp. 459–470, 2008.
- [7] 関洋平, 神門典子, 稲垣陽一, 栗山和子. 新聞記事とコミュニティ QA を対象とした詳細な意見分析コーパスの作成と分析. 情報処理学会研究報告, Vol. 2010-FI-97, pp. 1–8, 2010.
- [8] 中道龍三, 徳久雅人, 村上仁一, 池原悟. 情緒推定の手がかりとなる接続表現の収集. 電子情報通信学会技術研究報告, Vol. 108, pp. 1–6, 2008.