質問応答サイトを用いた意見テキストの収集と極性反転文の検出

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
‡ 東京工業大学大学院情報理工学研究科

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はなく，意見，評判，感想などの主観情報も存在する．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．

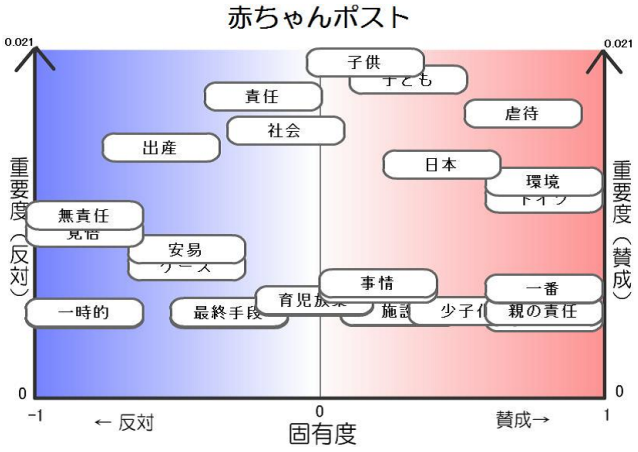

図 1: OpinionReader の出力例

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
(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」の機能改善を目的とする．

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，質問応答サイトから意見情報を収集する手法を提案する．質問応答サイトとは，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[1]や OKWave[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，質問応答サイトから「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．

[1] http://chiebukuro.yahoo.co.jp/
[2] http://okwave.jp/

Copyright © 2010 The Association for Natural Language Processing.
All Rights Reserved.

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もわかる。でもなんか納得がいかない。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。

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．

極性反転に関する研究として，那須川ら [5] や中道ら [7] 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．

3 提案する手法

3.1 概要

本研究は,「意見テキスト収集」と「極性反転文検出」のそれぞれに関する手法を提案する．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，質問応答サイトにおいて質問のタイトルに以下の文字列を含むページを検索する．

X に賛成 (？|?| 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．

(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 |[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」,「w」,「・」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」と「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| w |・)

ただし，表明を含む段落の抽出において以下の表現は立場の表明にはならない場合が多かったため，ストップワードとする．P は「賛成」または「反対」とする．

には P です

例えば,「賛成です。ただし、匿名性には反対です。」と書かれている場合には,「には反対です」がストップワードであるため,「賛成です。」という記述が規則に適合し，賛成意見として抽出する．

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．規則は極性が反転する表現を人手で分析し，作成した．極性が反転する表現を言語的特徴によって「否定」,「引用」,「譲歩」の 3 カテゴリに分類した．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，検出範囲が最小になるような規則を適用する．

以下，3.3.2〜3.3.4 で各カテゴリについて説明する．

Copyright © 2010 The Association for Natural Language Processing.
All Rights Reserved.

3.3.2 否定

「否定」に関する表現は，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．ここで,「が (助詞-接続助詞)」は ChaSen[3]による形態素解析を用いて判定する．

	手がかり表現	検出範囲
文頭	ただ, でも, しかし, ですが, けれども, だけど, けど, だが, もっとも, ただし	B
文中	が (助詞-接続助詞), としても, 反面, からといって, にせよ, けど	A
文末	が (助詞-接続助詞), けど, には X です	A

例えば，以下の例文における「ただ」という表現は，上記規則の「文頭」に該当する.「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．

例 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。

3.3.3 引用

「引用」に関する表現は，意見文中で他者の意見を引用する場合に使用される．例えば「などと言う」,「という意見」が表現となる．

この規則は，表現の後に続く単語が「言う」の類義語か,「意見」かによって 2 つのパターンに分けられる．各パターンにおける表現と検出範囲を以下に示す.

後に続く単語	手がかり表現	検出範囲
(i)「言う」の類義語	などと，だと，とか	A
(ii)「意見」	との，という	A

例えば，以下の例文における「とか言ってる」という表現は，上記規則の (i) に該当する．

例 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。

(i) の検出範囲は 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．

(ii) の規則は，例えば「育児放棄が増えるという意見がありますが」の「という意見」に該当する．

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」という表現が該当する．

例 反対です。確かに、赤ちゃんポストがあれば命は救われるでしょう。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．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．収集した情報のうち正解と判定したテキストは，賛成または反対の表明を含み，かつその立場の根拠が書かれているテキストである．また，表明に基づいて賛否に分類し，実際の意見と分類された立場が一致する場合のみを正解とした．

表 1 に結果を示す．表 1 の (d) と (f) における「意見数」とは，賛否の分類を含めて，正しく収集された意見情報の数である．表 1 の結果から，高い精度で賛成意見と反対意見を分類できることが分かった．

分類の誤りでは，表明のみで賛成または反対の根拠が書かれていない意見や，賛成，反対の両方の表明を含む中立意見があった．収集漏れは，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である．例えば，以下の表現は反対の立場を表明している．

例 赤ちゃんポストなんてやめて欲しい。

また，質問応答サイトにおける意見の偏りについて分析した．質問応答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 2 に「質問の影響量」と,「回答の影響量」の平均を示す.

表 2 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，質問応答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．

4.2 極性反転文検出の評価

極性反転文の検出手法について，Yahoo!知恵袋から人手で抽出した「赤ちゃんポスト」,「ゆとり教育」,「東京オリンピック」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 3 に示す.

[3]http://chasen.naist.jp/hiki/ChaSen/

Copyright © 2010 The Association for Natural Language Processing.
All Rights Reserved.

表 1: 意見収集手法の評価

	赤ちゃんポスト	ゆとり教育	東京オリンピック
(a) 検索されたページ数	19	8	34
(b) ページに含まれる回答数	160	64	182
(c) (b) のうち調査した回答数	160	64	100
(d) (c) のうち意見数	110	54	70
(e) 自動収集された段落数	121	46	64
(f) (d) のうち意見数	110	41	59
(g) 自動収集の精度	90.9%(110/121)	89.1%(41/46)	92.2%(59/64)
(h) 自動収集の再現率	83.3%(110/132)	75.9%(41/54)	84.3%(59/70)

表 2: 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 3: 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)

表 3 では，3 つのトピック全てにおいて 80% 以上の再現率が得られた．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，規則ごとの精度を分析した結果を表 4 に示す．

表 4: 規則ごとの検出精度

	(a) 否定	(b) 引用	(c) 譲歩
機械抽出	203	17	33
正解	95	15	21
精度	46.8% (95/203)	88.2% (15/17)	63.6% (21/33)

表 4 から，(a) 否定を表す規則の精度が特に低いことが分かった．「東京オリンピック」について，規則 (a) における誤り 41 件の内訳を表 5 に示す．

表 5: 「東京オリンピック」における規則 (a) の誤り内訳

全体の立場と検出箇所の立場が一致する	9 箇所
どちらの立場の根拠でもない	32 箇所

表 5 の結果より，誤りの多くは「どちらの立場の根拠でもない」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．

例 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，質問応答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．

参考文献

[1] Soo-Mim Kim and Eduard Hovy. Determing the sentiment of opinions. *Proceeding of Conference on Computational Linguistics*, pp. 1367–1373, 2004.

[2] Tomoyuki Nanno, Toshiaki Fujiki, Ysuhiro 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] 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.

Copyright © 2010 The Association for Natural Language Processing.
All Rights Reserved.