テキストマイニングにおける意見文の構造解析手法*

井上 結衣 (学籍番号 200821644)
研究指導教員: 石塚 英弘 (藤井 敦)

1. はじめに

World Wide Web 上の文書には，意見，評判，感想などの主観情報が含まれる．複数の人間が書いた主観情報から人々の考え方に対する傾向や法則を発見できれば，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．

Web 上の時事問題に対する意見テキストをマイニングし，その傾向を可視化するシステムとして OpinionReader [1, 4] がある．当システムでは，ある話題について賛否両論が対立する場合は「論点」が存在する点に着目し，賛成派と反対派が対立する構図を論点に基づいて 2 次元グラフ上に可視化する．論点とは，賛成また反対の根拠となる語句である．

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．

OpinionReader は，時事問題が与えられると「1. 意見収集」，「2. 論点抽出」，「3. 可視化」の処理を行う．しかし，可視化にとって重要な機能である「1. 意見収集」と「2. 論点抽出」に改善の余地がある．

意見収集では，時事問題に対する意見を Web から抽出し，賛否に基づいて分類する．従来の手法 [1] では，精度重視の「初期検索」と，網羅性重視の「再検索」の二段階検索を行う．しかし，精度と再現率のバランスを取ることが難しいという問題があった．

論点抽出では，1. で収集した意見情報から名詞句と動詞句を抽出して論点とする．例えば，意見全体の立場が「賛成」である意見情報から抽出した名詞句と動詞句を，賛成の論点とする．しかし，部分的に意見全体の立場とは異なる立場の主張を含んでいる場合がある．例えば，「赤ちゃんポストに賛成です。赤ちゃんの命が助かるからです。でもネーミングは悪いと思う。」という意見において，全体の立場は「賛成」である．しかし，「でもネーミングは悪

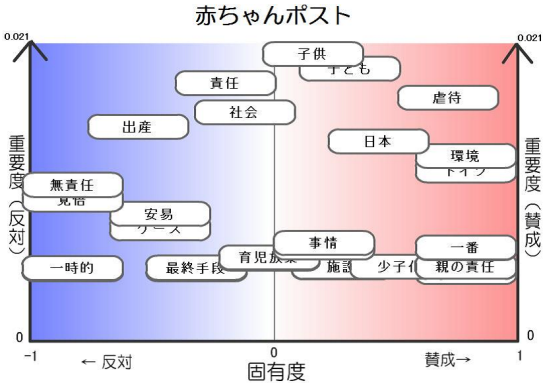

図 1　OpinionReader の出力例

いと思う。」という文は反対側の主張を述べている．本研究において，意見における著者の立場を「極性」と呼び，部分的に立場が反転している文を「極性反転文」と呼ぶ．上記の問題を解決するために，本研究は意見収集と極性反転文検出に関する手法を提案する．2 つの手法は，どちらも意見テキストの構造解析を行う点で共通している．

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や，文書中の主観的な記述を収集する手法がある [2]．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 [3] や中道ら [5] 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」などの逆接表現によって肯定や否定の極性が反転することを利用し，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した．しかし，どちらの研究も極性反転文検出を目的としていない点で本研究とは異なる．

3. 提案手法

3.1 概要

本手法は「意見収集」と「極性反転文検出」から成る．まず，意見収集は Web から意見情報を収集し，賛否に分類する．賛否に分類することによって意見全体の極性が得られる．次に，極性反転文検出によって，意見全体と極性が異なる文を検出する．

* “A Method for Analyzing Opinion Structures in Text Mining” by Yui INOUE

3.2 意見収集手法

本手法の特長は，QA サイトから意見収集を行い，賛否に基づいて分類する点にある．QA サイトとは，質問や回答の投稿によって知識を共有する Web サイトである．具体例として，Yahoo!知恵袋[†]や OKWave[‡]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．このようなページを，以降「QA ページ」と呼ぶ．ユーザは過去に投稿された QA ページを検索することもできる．

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 QA 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．

具体的には，まず質問タイトルに「入力された時事問題」と「賛成（？ |?| ですか | でしょうか)」を含む QA ページを検索する．次に検索された QA ページから「反対派です。」などの表明を含む回答を抽出し，表明に基づいて賛否に分類する．

3.3 極性反転文検出手法

手法を考案するために，意見テキストの実例を分析し，極性反転を言語の機能によって類型化した．その結果，以下に示す 3 つのカテゴリが得られた．

- 逆接：逆接表現によって態度を変える
- 引用：他者の意見を引用する
- 譲歩：逆の立場に譲歩する

各カテゴリに関する規則を人手で定義した．特徴的な表現を手がかり語句として定義し，手がかり表現ごとの検出範囲を定義した．例えば,「逆接」の手がかり表現として文中の接続助詞「が」を定義した．また，文中の「が」における表現ごとの検出範囲は「文頭から手がかり表現まで」と定義した．

手法を用いる際は意見テキストを文に分割し，規則を適用する．その結果, 条件に合致する箇所を「極性反転文」として検出する．例えば「赤ちゃんポスト賛成です。育児放棄は増えるが、命が助かる。」という文には,「逆接」の手がかり表現である「が」が含まれる．検出範囲は「文頭から手がかり表現まで」なので,「育児放棄は増えるが」の部分を極性反転文として検出する．

4. 評価実験

「赤ちゃんポスト」,「ゆとり教育」,「東京オリンピック」を対象のトピックとして，2 つの提案手法それぞれについて評価実験を行った．

意見収集の評価では, 本手法を用いて Yahoo!知恵袋から意見を収集し，賛否に分類した．収集したテキストについて人手で正解判定した．テキストが賛成または反対の根拠を含み，かつ正しく賛否に分類された場合に正解と判定した．3 トピックの合計で評価した結果，精度は 92.2 %，再現率は 83.2 %となり，精度と再現率共に高い値が得られた．

極性反転文検出手法の評価では，人手で収集した意見テキストに対して本手法を適用し，極性反転文の検出を行った．3 トピックの合計で評価した結果，再現率は 82.6 %と高い値が得られた．他方で，精度は 46.1 %と低い値になった．精度が低い原因として,「逆接」に関する規則における手がかり表現が逆接以外の多様な用法を持つ点がある．例えば,「逆接」の手がかり語句である文中の「が」は，逆接の他に「提示」や「打ち消し」の用法がある．用法が「逆接」の場合は極性反転する例が多い．しかし，それ以外の用法では極性は反転しないことが多い．今後は,「が」の用法を自動的に特定する必要がある．また，テキストマイニングのシステム全体における本手法の有効性を評価する必要がある．

文献

[1] 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.

[2] 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. 12, No. 2, pp. 111–118, 2006.

[3] 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.

[4] 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.

[5] 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.

[†] http://chiebukuro.yahoo.co.jp/

[‡] http://okwave.jp/