様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ５ 月 ２２ 日現在

機関番号：１３３０２
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００２９８

研究課題名（和文）他者の意見の引用を考慮したオピニオンマイニング

研究課題名（英文）Opinion Mining from Texts Including Quoted Others' Opinions

研究代表者
白井　清昭（Shirai, Kiyoaki）

北陸先端科学技術大学院大学・先端科学技術研究科・准教授

研究者番号：３０３０２９７０

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：社会問題や時事問題を対象としたオピニオンマイニングに向け，他者の意見を引用したテキストの極性を判定する手法を研究した．まず，与えられたテキストのうち他のテキストを引用している箇所を検出した．次に，ユーザが書いたテキストならびに他者が書いた引用箇所のテキストの極性を判定した．さらに，引用されたテキストと元のテキストの関係(順接，逆接，無関係)を推定した．これらを総合的に判断してテキスト全体の極性を決定した．評価実験の結果，提案手法による極性判定の正解率は0.942となり，引用を考慮しないベースライン手法の正解率0.893を大きく上回った．

研究成果の学術的意義や社会的意義
不特定多数のユーザが書いたテキストから特定の対象に対するユーザの意見や評判を明らかにするオピニオンマイニングは重要な研究課題である．特に時事問題を対象にしたオピニオンマイニングは，社会情勢を低コストで把握することができるため有用である．本研究課題は，他者の記事の引用を含むテキストの極性判定の精緻化によりオピニオンマイニングの正確性を向上させるものであり，その社会的意義は大きい．
これまでの極性判定の研究では，判定対象となるテキストの部分の性質の違いに着目した研究は少ない．本研究課題は，引用されたテキストとそうでないテキストを分けて処理することで極性判定の性能を向上させる点に学術的意義がある．

研究成果の概要（英文）：This research project aims at the polarity classification of texts (blog articles) including quotation of other articles toward opinion mining of social issues. First, a quoted text is extracted from a given blog article. Next, the polarity of a text written by a user and a quoted text written by others is identified. Then, a quotation type (“agree”, “disagree”, or “unrelated”) is identified, which stands for relation between the original and quoted texts. Finally, the polarity of the whole blog article is determined by considering the above results. In the experiments, the accuracy of the polarity classification of the proposed method was 0.942, which largely outperformed the baseline (0.893).

研究分野：自然言語処理

キーワード：オピニオンマイニング　極性判定　機械学習

１．研究開始当初の背景

近年では，ブログ，クチコミサイト，SNS などといった Consumer Generated Media (CGM) の普及に伴い，誰もが簡単に情報発信できる環境が整っている．これに伴ない，評判情報分析またはオピニオンマイニングと呼ばれる研究が盛んに行われている．オピニオンマイニングとは，主に CGM 上のテキストを分析し，特定の対象に対するユーザの意見や評判を明らかにする技術である．

オピニオンマイニングでは，その基本的な処理として，与えられたテキスト(文もしくは段落)に対し，それが肯定的もしくは否定的な意見を表明しているかを判定する．これはテキストの極性判定と呼ばれている．その際に有力な手がかりとなるのは評価語である．評価語とは，文が肯定的あるいは否定的かを示唆する単語(「よい」「素晴しい」「ダメ」「悪い」など)である．極性判定の基本的な原理は，評価語辞書，すなわち単語とその極性(肯定的か否定的か)の情報を含む単語のデータベースを利用し，テキストが肯定的もしくは否定的な単語のどちらを多く含むかで極性を判定する．もちろん，単純に評価語の数をカウントするだけでは肯定もしくは否定を正確に判定することは難しいため，機械学習の適用を始め様々な洗練された手法が提案されている．多くの CGM テキストを対象に極性判定を行い，その結果を集計することで，全体として肯定的な意見が多いか否定的な意見が多いかを知ることができる．

オピニオンマイニングの応用例の一つとして，社会問題や時事問題に対する世論を把握することが挙げられる．すなわち，ある社会問題に対して意見を述べているテキストを分析し，賛成意見・反対意見のどちらが多いかを調べる．また，社会問題に対する意見はブログで表明されることが多いため，分析の対象とする CGM としてはブログが適している．ところが，ブログでは，ユーザが新聞記事や他者のブログ記事を引用した上で自分の意見を述べることが多い．例えば，「プログラミングの必修化」という問題に対し，ユーザは好意的な新聞記事を引用した上で，それに対する反論を述べていることがある。記事の引用部分には肯定的な意見であることを示唆する評価語が多く含まれているため，従来のオピニオンマイニングの技術ではこのようなテキストに対して肯定的と誤判定する可能性が高い．このように，他のテキストの引用はオピニオンマイニングの正確性を低下させる原因になりうる．

２．研究の目的

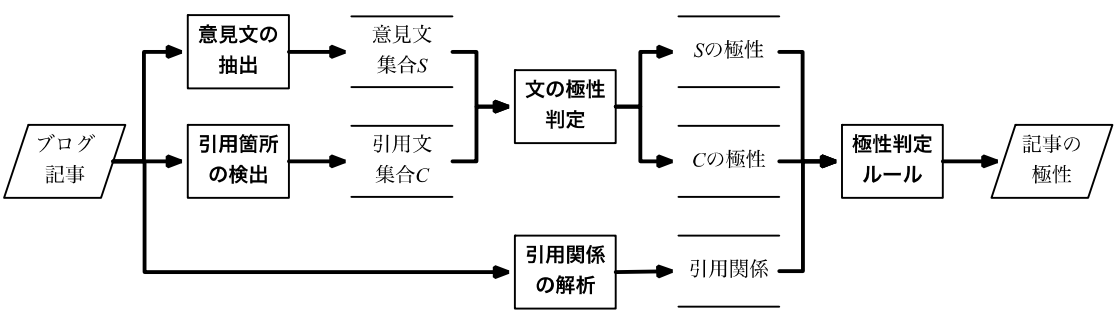
本研究課題は，与えられた時事問題に対し，それに対する意見を集約・分析・整理し，その全体像をわかりやすい形で提示する手法を確立することを目的とする．特に，テキスト間の引用・被引用関係に着目し，テキストの極性判定の精度を向上させることを狙う．

まず，ブログ記事が与えられたとき，他のテキストを引用している箇所を検出する技術を探究する．ブログでは，ユーザは実に様々な形式で他のテキストを引用している．引用を含む大量のテキストを網羅的に分析し，テキストの引用の特徴を明らかにする．さらに，その知見を基に，様々なブログ記事に対して引用箇所を正確に検出する汎用的な手法を開発する．

続いて，他のテキストの引用を含むテキストに対する極性判定の手法を確立する．ここでの極性判定とは，あるトピック(時事問題)に対し，テキストの書き手が表明している意見がトピックに対して賛成，反対，中立のいずれであるかを判定する処理を指す．まず，他のテキストの引用箇所以外のテキストの極性判定を行う．次に，引用されているテキストの極性判定を行う．さらに，元のテキストと引用されているテキストの関係を判定する．ここでの関係とは「順接」「逆接」などである．最後に，これらの解析結果を総合的に判断し，テキストの極性を判定する．例えば，引用されているテキストは賛成意見を表明していたとしても，元のテキストは逆接の関係でそのテキストを引用している場合には，極性を反転し，テキスト全体では反対意見を表明していると判定する．

３．研究の方法

あるトピックに関連するブログ記事を入力とし，そのブログ記事の極性，すなわち記事がトピックに対して賛成，反対，中立の意見を述べているのかを出力とするシステムを構築した．入力とするブログ記事には他者の記事の引用を含むと仮定する．提案システムの処理の流れは以下の図の通りである．



(1)意見文の抽出
ブログ記事の中から著者がトピックに対して意見を述べていると思われる文を抽出した．ここでは，トピックのキーワードを含む文，およびその前後 2 文を意見文として抽出した．以下，意見文の集合を S とおく．
(2)引用箇所の検出
ブログ記事から他の記事を引用している箇所を検出した．①引用を示唆するキーワード(転載，転載開始，引用，新聞，など)を含む DOM ノードもしくはそれに隣接している DOM ノード内の文を抽出する，②<hr>タグもしくは 4 つ以上同じ文字が続く文字列(テキスト境界を表すとみなせる)で囲まれた文を抽出する，③<blockquote>タグで囲まれた文を抽出する，などのルールにより引用箇所を特定した．以下，引用箇所に含まれる文の集合を C とおく．
(3)文の極性判定
S または C 中の文の極性が肯定，否定，中立のいずれかであるかを判定した．極性判定には機械学習による分類器を用いた．訓練データとして，筑波大学文単位評価極性タグ付きコーパス(以下，筑波コーパス)を用いた．同コーパスは楽天トラベルのレビューデータに対して文単位で評価極性情報を付与したコーパスである．同コーパスに付与された評価極性情報のうち，褒め(p)を「肯定」，苦情(k)を「否定」，ニュートラル(e)と要求(y)を「中立」とし，それ以外のタグが付与されている文は除去した．この結果，3 種類の極性タグが付与された 3757 文の訓練データが得られた．
機械学習に用いる素性は文中の自立語とした．また，自立語の後に否定表現が続く場合は否定を表すフラグを付与したものを素性とした．例えば，「嬉しくない」という文からは『嬉しい+否定』という素性を抽出した．素性の重みは，日本語極性評価語辞書に載っている語の重みは 2，それ以外の語の重みは 1 とした．極性判定の分類器として SVM を学習した．scikit-learn を使用し，カーネルは線形カーネル，正則化パラメタは 1.0 とした．
さらに，文集合 S の極性(Polarity(S)と記す)または C の極性(Polarity(C)と記す)を判定した．文集合 S のうち肯定と判定された文 s について，その SVM による判定の信頼度のスコアの和を肯定のスコアとした．否定のスコアも同様に計算した．そして，両者の差を S の極性スコアとし，これが閾値より大きいときは Polarity(S)を「肯定」，閾値より小さいときは「否定」，それ以外は「中立」と判定した．Polarity(C)も同様に決定した．
(4)引用関係の解析
本研究では，ブログ記事の著者がどのような立場で他のテキストを引用しているかを引用関係と呼ぶ．引用関係は，「順接」(引用したテキストの意見に対して賛成または同意しているとき)，「逆接」(引用したテキストの意見に対して反対しているとき)，「無関係」(単にテキストを引用しただけでそのテキストの意見に対する立場を表明していないとき)の 3 種とした．引用関係は，接続詞，動詞，引用関係を示唆する手がかり句などをもとに判定する．
(5)記事の極性判定
これまでの処理で決定された S の極性，C の極性，引用関係を手がかりに記事全体の極性を判定する 27 個の極性判定ルールを策定した．極性判定ルールの設計方針は以下の通りである．①S の極性はそのまま記事の極性とした．②C の極性が肯定または否定のとき，引用関係が順接ならそのまま，逆接なら反対の極性を記事の極性とした．③引用関係が無関係でかつ S の極性が肯定または否定のとき，S の極性を記事の極性とした．④引用関係が無関係でかつ S の極性が中立のとき，引用によって間接的に意見を表明していると判断し，C の極性を記事全体の極性とした．⑤S の極性で決まる記事の極性と，C の極性と引用関係で決まる記事の極性が矛盾するとき，タイブレークのルールを適用した．ここでタイブレークのルールとは，Polarity(S)と Polarity(C)の極性スコアの絶対値を比較し，それが大きい方の極性に決める処理である．

４．研究成果

提案手法の評価実験を行った．トピックとして「大阪都構想」「女系天皇」「夫婦別姓」の 3 つを選び，これに関するブログ記事を Yahoo!ブログから収集した．トピックを含み，かつそれに対する意見や他記事の引用を含むブログ記事を検索エンジンを用いて収集した．次に，収集したブログ記事を文に分割し，個々の文に対して，引用箇所か否か，文の極性(肯定，否定，中立)，引用関係の情報を人手でアノテーションした．さらに，記事全体の極性も人手で判定し，文の極性タグと同じ 3 種類のタグを付与した．これにより，およそ 300 記事，8,500 文からなる評価データを作成した．
提案手法による引用箇所検出の性能を評価したところ，精度が 0.791，再現率 0.641，F 値が 0.708 となった．3 つのトピックの比較では，「女系天皇」の F 値は 0.679 と一番低く，他の 2 つのトピックの F 値は 0.72 程度であった．比較的良好な結果が得られたと言える．一方，文単位の極性判定を評価したところ，極性判定の正解率は 0.653 となった．引用箇所検出よりも結果が悪いが，これは訓練データとして用いた筑波コーパスと今回の実験で用いた評価データのドメインの違いに起因すると考えられる．
次にブログ記事の極性判定手法を評価した．ここでは，引用箇所の抽出，文の極性判定，および引用関係の解析結果は人手でタグ付けした正解データを使用し，先に示した 27 個の極性判定ルールを用いて記事の極性を決定した結果を評価した．引用を考慮せずにブログ記事の全ての文を対象に記事の極性を決める方法をベースラインとした．このベースラインの正解率が 0.893

であるのに対し，他者の記事の引用を考慮した提案手法の正解率は 0.942 となった．正解率がおよそ 0.05 ポイント向上したことから，引用関係を考慮して記事の極性を判定する提案手法の有効性が確認された．3 つのトピック毎に提案手法の正解率を比較すると，「大阪都構想」が 0.952(+0.013)，「女系天皇」が 0.902(+0.024)，「夫婦別姓」が 0.961(+0.049)であった．括弧内はベースラインとの差であり，特に「夫婦別姓」のトピックについてベースラインとの差が大きかった．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Hy Nguyen, Kiyoaki Shirai
２．発表標題 A Joint Model of Term Extraction and Polarity Classification for Aspect-based Sentiment Analysis
３．学会等名 The 10th International Conference on Knowledge and Systems Engineering. pp.323-328（国際学会）
４．発表年 2018年

１．発表者名 村松健太, 白井清昭
２．発表標題 引用関係の解析に基づくテキストの極性判定
３．学会等名 言語処理学会第25回年次大会. pp.1149-1152
４．発表年 2019年

１．発表者名 Kiyoaki Shirai, Yunmin Xiang
２．発表標題 Over-sampling Methods for Polarity Classification of Imbalanced Microblog Texts
３．学会等名 The 33th Pacific Asia Conference on Language, Information and Computation（国際学会）
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考