引用関係の解析に基づくテキストの極性判定

村松 健太　　白井 清昭
北陸先端科学技術大学院大学 先端科学技術研究科
{s1710201,kshirai}@jaist.ac.jp

1 はじめに

オピニオンマイニングは，トピックに対する人々の様々な意見を解析・集約し，これらを俯瞰的に見るための情報を提供する技術である [1, 4]．オピニオンマイニングに必要とされる処理のひとつに，テキストがトピックに対して肯定的か否定的かを判定する極性判定が挙げられる．ウェブ上のテキスト，特にブログ記事では，他者の記事を引用した上で意見が述べられることがあるが，このようなテキストに対する極性判定は注意を要する．例えば，肯定的な意見を引用した上で,「この意見には賛成できない」といった表現で否定的な見解を述べる場合があるが，引用箇所には肯定的な評価語が出現するため，テキストの極性を肯定的と誤って判定する可能性がある．

岡山と白井は，引用を含むブログ記事に対し，引用箇所を検出し，それを除去した上でテキストの極性を判定する手法を提案した [3]．しかし，彼らの手法では単に引用箇所を除去するだけであり，書き手がどのような立場で他のテキストを引用しているかは考慮されていない．テキスト間の関係を考慮した極性判定として，Zhou らは，Twitter の極性を判定する際にフォロー関係を機械学習の素性として利用する方法を提案している [5]．しかし，引用された他者の記事と元のテキストの関係に着目しているわけではない．

本論文は，引用を含むブログ記事に対し，引用されたテキストとブログ著者の意見の関係を解析し，その結果を利用して極性判定の正解率を高める手法を提案する [2]．

2 提案手法

提案手法における処理の流れを図 1 に示す．あるトピック (「大阪都構想」など) 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．以下，各モジュールの詳細について説明する．なお，前処理として，ブログ記事はあらかじめ文に分割しておく．

2.1 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する．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する．以下，意見文の集合を S とおく．

2.2 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する．岡山と白井の方法 [3] にならい，以下のルールを用いて引用箇所を検出する．

- 引用を示唆するキーワード[1]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
- ⟨hr⟩ タグもしくは 4 つ以上同じ文字が続く文字列[2]で囲まれた文
- ⟨blockquote⟩ タグで囲まれた文

以下，引用箇所に含まれる文の集合を C とおく．

2.3 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する．極性判定には機械学習による分類器を用いる．訓練データとして，与えられたトピックに対して意見を表明し，その極性が付与されている文の集合を用いることが理想的であるが，実際の応用では様々なトピックが入力として与えられることから，このようなデータを事前に用意することは難しい．そこで，筑波大学文単位評価極性タグ付きコーパス[3](以下，筑波コーパス) を訓練データとして用いる．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．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する．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．

[1]転載，転載開始，引用，毎日新聞，など
[2]テキスト境界を表すとみなせる．例えば「＊＊＊＊＊」など．
[3]http://www.nlp.mibel.cs.tsukuba.ac.jp/~inui/SA/corpus/

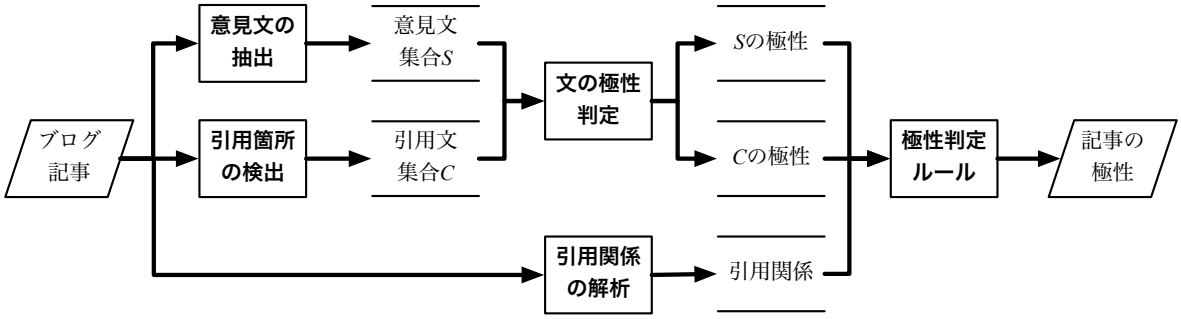

図 1: 提案手法の概要

機械学習に用いる素性は文中の自立語とする．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する．例えば,「嬉しくない」という文からは『嬉しい+否定』という素性を抽出する．素性の重みは，日本語極性評価語辞書 (用言編) と (名詞編) に載っている語の重みは 2，それ以外の語の重みは 1 とする．筑波コーパスから抽出された素性数は 1,373 であった．

極性判定の分類器として SVM を学習する．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 とする．

さらに，文集合 S または C の極性 (Polarity(S) または Polarity(C) と記す) を判定する．まず，式 (1) のように，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 $svm\text{-}score(s)$ の和を肯定のスコア $\mathrm{S}_{pos}(S)$ とする．否定のスコア $\mathrm{S}_{neg}(S)$ も同様に計算する．両者の差を S の極性スコア $\mathrm{S}_{POL}(S)$ とする (式 (2))．最後に，式 (3) に示すように，$\mathrm{S}_{POL}(S)$ が十分大きいとき (閾値 d より大きいとき) には Polarity(S) を肯定，十分小さいときには否定，絶対値が大きくないときは中立と判定する．Polarity(C) も同様に判定する．

$$\mathrm{S}_{pos}(S) = \sum_{s \in S \wedge pol(s)=\text{肯定}} svm\text{-}score(s) \quad (1)$$

$$\mathrm{S}_{POL}(S) = \mathrm{S}_{pos}(S) - \mathrm{S}_{neg}(S) \quad (2)$$

$$\mathrm{Polarity}(S) = \begin{cases} \text{肯定} & \text{if } \mathrm{S}_{POL}(S) > d \\ \text{否定} & \text{if } \mathrm{S}_{POL}(S) < -d \\ \text{中立} & \text{if } |\mathrm{S}_{POL}(S)| \leq d \end{cases} \quad (3)$$

2.4 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．引用関係は以下のいずれかと定義する．

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．

引用関係を解析する方法については現在検討中である．およそ以下の方針で引用関係を決定することを考えている．

- 接続詞による判定.「しかし」「ところが」など逆接の接続詞があれば引用関係を逆接と判定する．
- 引用関係を示唆する手がかり句による判定．例えば,「正論」「妥当」は順接,「的はずれ」「間違っている」は逆接,「ちなみに」「さて」は無関係を示唆すると考えられる．

2.5 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，ルールベースの手法で記事全体の極性を判定する．極性判定ルールの詳細を表 1 に示す．TB は後述のタイブレークのルールで記事の極性を決定することを表す．これらのルールは以下の方針にしたがって設計されている．

- S の極性はそのまま記事の極性とする．
- C 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する．
- 引用関係が無関係のとき，
  - S の極性が肯定または否定のとき，それを記事の極性とする．(ルール R3, R6, R9, R12, R15, R18)
  - 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する．(ルール R21, R24, R27)

表 1: 極性判定ルール

	S の極性	C の極性	引用関係		記事の極性
R1	肯定	肯定	順接	→	肯定
R2	肯定	肯定	逆接	→	TB
R3	肯定	肯定	無関係	→	肯定
R4	肯定	否定	順接	→	TB
R5	肯定	否定	逆接	→	肯定
R6	肯定	否定	無関係	→	肯定
R7	肯定	中立	順接	→	肯定
R8	肯定	中立	逆接	→	肯定
R9	肯定	中立	無関係	→	肯定
R10	否定	肯定	順接	→	TB
R11	否定	肯定	逆接	→	否定
R12	否定	肯定	無関係	→	否定
R13	否定	否定	順接	→	否定
R14	否定	否定	逆接	→	TB
R15	否定	否定	無関係	→	否定
R16	否定	中立	順接	→	否定
R17	否定	中立	逆接	→	否定
R18	否定	中立	無関係	→	否定
R19	中立	肯定	順接	→	肯定
R20	中立	肯定	逆接	→	否定
R21	中立	肯定	無関係	→	肯定
R22	中立	否定	順接	→	否定
R23	中立	否定	逆接	→	肯定
R24	中立	否定	無関係	→	否定
R25	中立	中立	順接	→	中立
R26	中立	中立	逆接	→	中立
R27	中立	中立	無関係	→	中立

- S 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，以下のタイブレークのルールを適用する．

  タイブレーク

  $\mathrm{S}_{POL}(S)$ と $\mathrm{S}_{POL}(C)$ の絶対値を比較し，大きい方の極性に決める．同じときは中立と判定する．

3 評価実験

3.1 評価データの作成

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[4]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収集するため，以下の検索クエリを用いた．

> [トピック] (賛成 OR 反対 OR 新聞 OR ニュース OR 引用 OR 抜粋 OR 掲載 OR 転載) site:blogs.yahoo.co.jp

検索エンジンは Goole 検索を用いた．「新聞」「ニュース」などのキーワードを用いているのは，予備調査では新聞記事が引用されることが多かったためである．

次に，収集したブログ記事を文に分割し，個々の文に対して以下の情報をアノテーションした．アノテーションは第 1 著者が実施した．

- 引用タグ = { 引用，引用ではない }
- 極性タグ = { 肯定，否定，中立 }
- トピックタグ = { トピックに言及している，していない }
- 引用関係 = { 順接，逆接，無関係 (言及あり)，無関係 (言及なし) }

  引用について言及しているがそれに対する自身の立場を表明していない文は「無関係 (言及あり)」，引用について言及していない文は「無関係 (言及なし)」とした[5]．

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．作成した評価データおよびアノテーションの詳細を表 2 に示す．

3.2 実験結果

引用箇所検出の実験結果を表 3 に示す．評価基準は「引用」とタグ付けされた文の精度，再現率，F 値である．F 値は 0.7 程度と比較的良好な結果が得られた．

次に文単位の極性判定手法を評価する．評価基準は，正解率 (予測した極性タグが正解と一致した文の割合) ならびに肯定，否定タグを持つ文の F 値とした．参考として，筑波コーパスの 5 分割交差検定による極性判定の結果も評価した．結果を表 4 に示す．評価データにおける肯定クラス，否定クラスに対する F 値は低く，実用的に十分とは言えない．F 値と比べて正解率は 0.58〜0.73 程度と高いが，これは評価データの極性タグが中立に偏っているためと考えられる．また，評価データの極性判定の F 値は筑波コーパスよりもかなり低い．これは訓練データと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する．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，表 1 に示した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する．すなわち，引用関係を解析することで極性判定の正解率がどれだけ向上するかを検証することを目的とする．以下の 3 つの手法を比較する．

ベースライン (BL) 引用を考慮せずに極性を判定する．ブログ記事の全ての文を対象に式 (3) にし

[4] https://blogs.yahoo.co.jp/

[5] 提案システムでは両者はともに「無関係」として扱う．

表 2: 評価データ

	記事数	文数	引用タグ		極性タグ			引用関係			
			あり	なし	肯定	否定	中立	順接	逆接	UR1	UR2
大阪都構想	100	3661	676	2985	54	280	3327	16	16	20	3690
女系天皇	100	1764	276	1488	32	193	1540	11	8	30	1715
夫婦別姓	100	3491	697	2794	51	246	3194	19	13	26	3433

UR1 = 無関係 (言及あり)，UR2 = 無関係 (言及なし)

表 3: 引用箇所検出の評価結果

	精度	再現率	F 値
大阪都構想	0.815	0.649	0.722
女系天皇	0.761	0.613	0.679
夫婦別姓	0.799	0.662	0.724
(全て)	0.791	0.641	0.708

表 4: 文の極性判定の評価結果

	正解率	F 値 (肯定)	F 値 (否定)
筑波コーパス	0.646	0.744	0.472
大阪都構想	0.731	0.0424	0.137
女系天皇	0.577	0.0197	0.118
夫婦別姓	0.683	0.0341	0.122
(全て)	0.653	0.0276	0.121

たがって極性を決める．文の極性判定の信頼度 $svm\text{-}score(s)$ は常に 1，閾値 d は 0 とした．

引用関係を解析しない手法 (P_{-C}) 記事の引用を考慮するが，引用関係は考慮しない手法．引用箇所を無視し，S の極性によって記事の極性を決定する．基本的な考え方は文献 [3] と同じである．

提案手法 (P_{+C}) 引用関係を考慮した極性判定ルール (表 1) によって記事の極性を決定する．

各手法の記事単位の極性判定の正解率，肯定クラスの F 値，否定クラスの F 値を表 5 に示す．提案手法 P_{+C} は，例外的に女系天皇のトピックでは P_{-C} より正解率や肯定クラスの F 値が低いが，全体的には BL や P_{-C} を上回っている．この結果から，引用関係を考慮してブログ記事の極性を判定する提案手法の妥当性が確認された．

表 5: ブログ記事の極性判定の評価結果

トピック	手法	正解率	F 値 (肯定)	F 値 (否定)
大阪都構想	BL	0.939	0.957	0.898
	P_{-C}	0.916	1.000	0.837
	P_{+C}	0.952	1.000	0.920
女系天皇	BL	0.878	0.400	0.909
	P_{-C}	0.927	1.000	0.909
	P_{+C}	0.902	0.500	0.941
夫婦別姓	BL	0.902	0.846	0.929
	P_{-C}	0.882	0.870	0.929
	P_{+C}	0.961	0.956	0.983
(全て)	BL	0.893	0.775	0.912
	P_{-C}	0.888	0.935	0.893
	P_{+C}	0.942	0.938	0.950

4 おわりに

本論文は，他者のテキストを引用しつつ自身の見解を述べているテキストに対し，引用箇所の極性判定および引用関係の解析によって記事単位の極性判定の正解率を向上させる手法を提案した．評価実験の結果，本研究の構想の有効性を確認した．今後の課題として，未実装となっている引用解析の手法の実現，文単位の極性判定の性能向上，より大規模な評価データによる提案手法の評価，などが挙げられる．

参考文献

[1] 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 D, Vol. J91-D, No. 2, pp. 459–470, 2008.

[2] 村松健太. 引用関係の解析に基づくテキストの極性判定. 修士論文, 北陸先端科学技術大学院大学, 3 2019.

[3] 岡山有希, 白井清昭. 他者のコメントの引用を考慮したオピニオンマイニング. 言語処理学会第 19 回年次大会, pp. 814–817, 2013.

[4] Hideyuki Shibuki, Takahiro Nagai, Masahiro Nakano, Rintaro Miyazaki, Madoka Ishioroshi, and Tatsunori Mori. A method for automatically generating a mediatory summary to verify credibility of information on the Web. In *Proceedings of COLING*, pp. 1140–1148, 2010.

[5] Xujuan Zhou, Enrico Coiera, Guy Tsafnat, Diana Arachi, Mei-Sing Ong, and Adam G. Dunn. Using social connection information to improve opinion mining: Identifying negative sentiment about HPV vaccines on Twitter. In *Proceedings of MEDINFO*, pp. 761–765, 2015.