

# 周辺語と構文的特徴に着目した Web 議論掲示板の 固有表現に対する感情極性判定

安藤 智大<sup>1,a)</sup> 藤田 桂英<sup>1,b)</sup>

**概要:** Web 議論掲示板において、投稿や文章に対し感情極性を判定する手法が数多く提案されている。しかし、複数の論点が同時に展開される大規模な Web タウンミーティングでは、出現するキーワードごとに極性を付与させる手法が新たに必要となる。そこで本稿では、Web 議論掲示板において出現した固有表現に対し、発言中において肯定的であるか、否定的であるか、あるいはそのどちらでもないかを判別する手法を提案する。提案手法では、評価極性辞書と Web 議論コーパスを用いた教師あり学習を行うことで自動分類する。学習には固有表現の周辺語から抽出する特徴を 2 つ、構文的特徴を 1 つ定義し使用する。その後、5 人の学生によって作成された正解データと比較することで評価を実施する。評価実験において、周辺語特徴が F 値においてベースラインを上回ることが確認され、構文的特徴は一部の場面でのみ効果を得られることが明らかになった。

## 1. はじめに

近年、地方の自治体では、地域住民の意見を集約し政策に反映させるタウンミーティングを、Web 上で実施する試みが進められている。対面式の議論とは異なり、大人数がオンライン上で議論し意見交換を行うことで、時間的・空間的な制約にとらわれないという利点を持つ。例として、大規模議論システム COLLAGREE では、Web 上におけるタウンミーティングの実施、参加者支援機能の開発等を行っている [1]。このような Web 議論掲示板では、参加者の発言がある論点に対し肯定的であるか、否定的であるかを明確にする感情極性判定が重要である。1 つの議題に対し賛成派、反対派に分かれる一般的な議論と異なり、Web 上におけるタウンミーティングでは、議論が進行するにつれ複数の議題・論点が出現し、また 1 つの投稿中に複数の意見を述べる場合も多い。そのため、各議題・論点ごとに感情極性を付与させる必要が生じる。

本論文では、Web 議論掲示板における各投稿に出現する固有表現に着目し、それぞれの固有表現に対して、ポジティブ、ネガティブ、中立(感情極性無し)の 3 種類のいずれかに自動分類することを目的とする。

本論文の構成を述べる。第 2 章では、関連研究として、感情極性分析に関する詳細を示す。第 3 章と第 4 章では、提案手法の詳細として、評価極性辞書の自動拡張と、その辞書を利用した教師あり学習について述べる。第 5 章では実験の結果と考察を、第 6 章で本論文のまとめを示す。

## 2. 関連研究

近年、感情極性を分析する手法は多種の Web サービスに応用されており、その応用先の特性を考慮した手法の提案が課題となる。2002 年に Pang らは、映画レビューサイトの投稿に対し、ナイーブベイズ、最大エントロピー原理、サポートベクターマシン (SVM) による教師あり学習によるポジティブ・ネガティブの分類手法を提案し、比較実験をした [2]。以降、機械学習を用いた分類手法が数多く提案され、特に SNS やマイクロブログの急速な発展に伴い、Twitter の呟きに対する分析手法が多く提案されている。例えば、Soroush らは、Twitter の呟きデータに対し、前後の呟きから文脈情報を取得し学習する手法を提案した [3]。また、近年では大規模コーパスを用いた深層学習による手法も多く存在する。Chen らは、商品レビューサイトの評価値(☆1 から☆5 までの 5 段階)の自動分類という課題に対し、深層学習モデルの一種である LSTM モデルと、ユーザ情報及び商品情報によるアテンションモデルを組み合わせた手法を提案した [4]。LSTM モデルとアテンションモ

<sup>1</sup> 東京農工大学 工学部 情報工学科  
Tokyo University of Agriculture and Technology

a) ando@katfujilab.tuat.ac.jp

b) katfujii@cc.tuat.ac.jp

表 1 分類ラベルと定義

ラベル	定義
ポジティブ	肯定, 賛同, 奨励, その他議論中でプラスの意味で使用されている語句
ネガティブ	否定, 反対, 不満, 疑問, その他議論中でマイナスの意味で使用されている語句
中立	単なる情報として与えられていて, 文脈上感情的でない語句

デルによる深層学習は, 従来の再帰ニューラルネットワークと比較し, 系列データに有用であることが確認されているため, 近年の自然言語処理において多く使用されているモデルである. Zhou らは, 異なる言語に対する感情極性分析をこれらのモデルで行うなど, 応用先が幅広い [5].

感情極性分析は, 分析するデータの単位によっても分類手法が異なる. 分析データの単位は大別して3つに分類され, 文書単位, 文章単位, 語句単位の順でより詳細な分類となる. 語句単位による感情極性の自動分類として, Wilson らは, 文章中に現れる語句の感情極性は, 文脈に依存して決定されると定義付け, 文中の句に対する特徴素を設計し学習・分類を行った [6].

また, 近年 Web 議論掲示板における自動分類や自動抽出に関する研究が多数行われている中で, 議論参加者の賛成・反対に関する立場を自動的に明確にすることは極めて重要であり, 参加者のスタンスを分析する手法がこれまでに多数提案されてきた [7][8].

### 3. 評価極性辞書の自動拡張

従来の評価極性辞書の語彙数では大規模議論実験において効果を発揮することが難しい. 本論文では, 既存の評価極性辞書をシードデータとした辞書の自動拡張を行い, 汎用的評価極性辞書を作成する. 既存の評価極性辞書として, 日本語評価極性辞書 (用言編) 及び日本語評価極性辞書 (名詞編)\*1を統合した辞書 (既存辞書) を用いる. 日本語評価極性辞書は, 東北大学の乾・岡崎研究室が公開リソースとして公表している評価表現リストである. 用言編は, 小林らが収集した用言を中心とした評価表現約 5,000 件のリスト [9] に一部改編を加え, 人手による評価極性情報を付与したデータである. 名詞編は, 東山らの手法により獲得した評価極性を持つ名詞及び複合名詞合わせて約 8,500 件に対し, 評価極性情報の人手によるチェックを行ったデータである [10].

辞書拡張の手法は, 日本語大規模データを用いて実装する. 以下に, 既存辞書を拡張するまでの手法の流れを示す.

- (1) Wikipedia の日本語記事データ\*2を学習し, 単語の分散表現を取得する.
- (2) 既存辞書に存在する単語を 1 で得た分散表現に変換する.
- (3) 2 で取得した分散表現とその語の極性から分類器を作

\*1 <http://www.cl.ecei.tohoku.ac.jp/>

\*2 <https://dumps.wikimedia.org/jawiki/latest/>

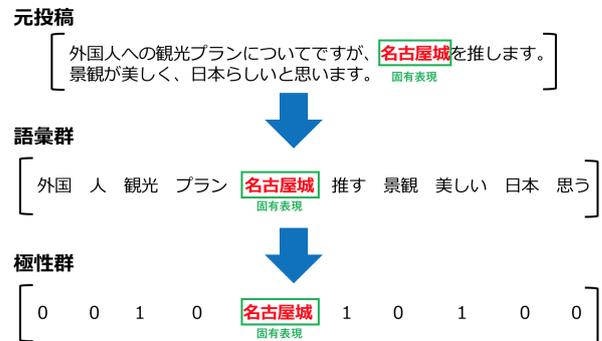


図 1 語彙群, 極性群の獲得の例

成する.

- (4) 辞書に存在しない語句の分散表現から極性を分類する.

提案手法では, 構築したベクトル空間モデルにおいて, 既存辞書に存在する単語を検索し, 分散表現を取得する. なお分散表現の取得には Word2Vec を用い, 200 次元のベクトル空間モデルを構築した. 取得した各分散表現を特徴素とし, 対応する極性, すなわちポジティブ, ネガティブ, 中立のいずれかをラベルとする. 特徴素とラベルのセットを学習データとして, SVM を用いた 3 値分類学習を行う. 本論文では, 自動拡張の対象とする単語を, 既存辞書に存在せず, かつ日本語記事データ中に 10 回以上出現している単語とする. 学習により作成した分類器に, 対象単語の分散表現を当てはめることでそれらの感情極性を自動で判定する.

### 4. 議論中の固有表現に対する感情極性判定手法

本論文では, 大規模議論掲示板において, 投稿中に出現する固有表現を対象とした 3 種のラベルへの分類を, 教師あり学習によって行う. 表 1 に各ラベルの分類区分を示す. 学習及び分類には, 投稿文中の固有表現ごとに 3 つの素性を新たに定義し用いる.

#### 4.1 各投稿に対する群の定義

特徴素取得の前処理として, 各投稿文から体言及び用言以外の形態素を削除した群 (語彙群) を定義する. なお用言は原型に変換する. また, 語彙群の各語に対し評価極性辞書を参照し, 参照先のラベルに変換した群 (極性群) を定義する. 図 1 は, 投稿文から語彙群, 極性群を獲得する例を示す. 極性群の数字は, ポジティブ:1, ネガティブ:-1, 中立:0 にそれぞれラベルが対応している.

## 4.2 素性群 $\alpha$ :語彙 N-gram

語彙群において出現した固有表現に対し、自身を含む周辺語 N-gram を抽出し、語彙の出現回数の合計を特徴化する。例として、図1の語彙群において  $N=4$  とすると、(人, 観光, プラン), (観光, プラン, 推す), (プラン, 推す, 景観), (推す, 景観, 美しい) の4つの語彙集合を得る。全ての語彙集合から語彙の出現回数を記録し、辞書化を行う。例の場合は、[外国:0, 人:1, 観光:2, プラン:3, 推す:3, 景観:2, 美しい:1, 日本:0, 思う:0] という辞書が作成される。全ての固有表現に対し周辺語彙の辞書化を実施し、各次元を語彙の出現回数としたベクトルを作成し  $\alpha$  とする。

## 4.3 素性群 $\beta$ :極性 N-gram

語彙 N-gram と同様の手順を、極性群に対し実行する。図1の極性群において  $N=4$  とすると、(0, 1, 0), (1, 0, 1), (0, 1, 0), (1, 0, 1) の4つの極性集合を得る。全ての極性集合から極性の出現回数を記録し、[ポジティブ, ネガティブ, 中立] の順でベクトル化する。例の場合は、[6, 0, 6] というベクトルが作成される。作成したベクトルに対し、式1で表される標準化を行った3次元のベクトルを素性群  $\beta$  とする。式1において、 $X$  は各極性の出現回数、 $\mu$  は出現回数の平均値、 $\sigma$  は標準偏差であり、 $z$  は平均0, 分散1の標準正規分布に従う。

$$z = \frac{X - \mu}{\sigma} \quad (1)$$

## 4.4 素性群 $\gamma$ :係り受け極性

素性群  $\beta$  は、対象の固有表現の周辺語から感情極性を読み取る効果があるが、周辺語には感情に起因しない語句が含まれている場合が多い。対象の固有表現とより密接な共起関係にあたる語を検索する手法として、係り受け解析を用いる。例として「名古屋城は鯨が目印の天守閣が魅力です」という文章に対し、固有表現である「名古屋城」が含まれる係り受け文節ペアを取得すると、(名古屋城は, 魅力です) というペアが得られる。固有表現と係り受けの関係になっている文節の体言及び用言を係り受け語群として定義する。先述したペアの場合「魅力」が語群として追加される。係り受け語を全て係り受け語群に追加し終えたら、極性群と同様に極性辞書を用いた変換、出現回数の記録、標準化を行い3次元のベクトルに変換する。変換したベクトルを素性群  $\gamma$  とする。

## 4.5 Random Forest を用いた学習と分類

提案手法では素性群  $\alpha$ ,  $\beta$ ,  $\gamma$  を組み合わせ学習及び分類を行う。分類器には Random Forest を用いる。

## 5. 評価実験

### 5.1 データセット

学習データ及び正解データとして扱う感情極性付き固有表現コーパスを、学生5人のアノテーションによって作成した。コーパス作成にあたり使用した議論掲示板データセットの詳細とアノテーション結果を表2に示す。議論データは全て Web 議論掲示板 COLLAGREE より引用したものであり、mecab-ipadic-NEologd の形態素解析により「固有名詞」と判定された語句を議論中の固有表現として抽出した。また、不適と判定されたデータは本実験で使用しない。

### 5.2 実験設定

提案した素性群を組み合わせ、感情極性付き固有表現コーパスを対象に8分割交差検証を行い、適合率、再現率、F値による評価を行う。ベースラインとして、対象の固有表現が含まれる投稿文中の単語数を素性とした Bag of Words(BoW)を用いる。また、辞書の自動拡張について比較を行うために、素性群  $\beta$  と素性群  $\gamma$  を既存辞書のみで作成する手法も評価を行う。実験は下記の2つの場合に分けて実施する。

- I. ポジティブ, ネガティブ, 中立の3値分類
  - II. ポジティブ, ネガティブのそれぞれに対する2値分類
- 実験 I では、ポジティブを1, ネガティブを-1, 中立を0とした3つのラベルに対して、RandomForestによる学習及び分類を実施する。実験 II では、ポジティブを1, それ以外を0とする2値分類器, ネガティブを1, それ以外を0とする2値分類器の2つを作成しそれぞれに対して議論データセットから分類を実行する。

提案手法におけるパラメータの設定について記す。素性群  $\alpha$ , 素性群  $\beta$  は、周辺語の範囲である  $N$  の値によって取得される語数、極性値が変化する。評価実験ではどちらも  $N=20$  とした。また、本実験で用いるデータセットは、中立ラベルが非常に多く、またネガティブラベルが極端に少ない不均衡コーパスである。したがって、各実験では、学習の際に教師データのアンダーサンプリング及びオーバーサンプリングを実行する。実験 I では全てのラベルについて、ポジティブの個数を基準値とした。また実験 II では、学習時のラベルの値が1である素性の個数を算出し、その2倍の数を基準値とする。ラベルの値が1の素性と0の素性の両方が基準値になるようにアンダーサンプリング及びオーバーサンプリングを実行する。アンダーサンプリングはランダム選択, オーバーサンプリングは SMOTE アルゴリズムを用いてそれぞれ調整を行った。

表 2 データセットとアノテーション結果

議論テーマ	投稿者数	投稿数	固有表現数	ポジティブ	ネガティブ	中立	不適
名古屋の魅力について語ろう	827	1351	8120	1238	349	5664	869
名古屋における環境	20	261	1153	46	43	743	321
名古屋における災害	21	332	1411	44	86	823	458

表 3 実験 I の評価結果

手法	ポジティブ			ネガティブ			中立		
	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
BoW	0.275	0.658	0.388	0.190	0.385	0.252	0.901	0.608	0.726
$\alpha$	0.294	0.718	0.417	0.279	0.423	0.334	0.907	0.635	0.747
$\alpha + \beta$ (辞書拡張なし)	0.301	0.718	0.424	<b>0.304</b>	0.448	<b>0.356</b>	0.907	0.646	0.754
$\alpha + \beta + \gamma$ (辞書拡張なし)	<b>0.307</b>	<b>0.725</b>	<b>0.432</b>	0.291	<b>0.450</b>	0.350	<b>0.910</b>	0.649	0.758
$\alpha + \beta$ (辞書拡張あり)	0.304	0.705	0.424	<b>0.304</b>	0.411	0.348	0.904	<b>0.663</b>	<b>0.765</b>
$\alpha + \beta + \gamma$ (辞書拡張あり)	0.298	0.711	0.420	<b>0.304</b>	0.391	0.339	0.904	0.658	0.762

表 4 実験 II の評価結果

手法	ポジティブ			ネガティブ		
	適合率	再現率	F 値	適合率	再現率	F 値
BoW	0.316	0.542	0.399	0.156	0.508	0.239
$\alpha$	0.356	0.602	0.447	0.215	0.532	0.306
$\alpha + \beta$ (辞書拡張なし)	0.382	<b>0.614</b>	0.471	0.219	<b>0.578</b>	<b>0.317</b>
$\alpha + \beta + \gamma$ (辞書拡張なし)	<b>0.390</b>	<b>0.614</b>	<b>0.477</b>	0.217	0.545	0.310
$\alpha + \beta$ (辞書拡張あり)	0.372	0.596	0.458	<b>0.221</b>	0.539	0.310
$\alpha + \beta + \gamma$ (辞書拡張あり)	0.376	0.592	0.460	0.215	0.513	0.302

### 5.3 実験結果と考察

表 3 と表 4 はそれぞれ実験 I と実験 II の評価結果を示す。全ての実験、ラベルにおいてベースラインである Bag of Words を提案手法が適合率、再現率、F 値ともに上回り、また素性群  $\beta$  を追加することによる精度の上昇が確認された。一方で、ポジティブ以外のラベルでは係り受け極性  $\gamma$  を追加しても、精度の上昇に繋がらないことが確認され、また辞書拡張が中立ラベル以外で効果を発揮していないことが確認された。

係り受け極性が効果を発揮しない理由として、使用コーパスの多くが形式的な日本語の文章でないことが挙げられる。例として「A はいいですね。もちろん B も」のような付け足しの文章や、「私はおすすめしたいですね。C」のように句点によって適切に係り受け解析ができない文章がある。このようなくだけた日本語に対し係り受け解析の効果は期待できないため、不要なデータの抽出が増加する。

また、辞書拡張が効果を発揮しない理由として、拡張辞書に不要な語が多いことが挙げられる。拡張辞書のヒット回数上位には、「ない」「ある」「思う」「する」「いる」など、汎用的に使用される語が多く、極性判定としては不要な語である。「少ない」「残る」など感情極性に関連する語もいくつか存在するため、人手による評価での語句選択や、登場回数が極端に多い語句の自動除去など、更なる改善が必要である。

## 6. おわりに

本論文では、Web 議論掲示板における各投稿に出現する固有表現に着目し、それぞれの固有表現に対して、ポジティブ、ネガティブ、中立の 3 種類のいずれかに自動分類することを目的とし、各固有表現の特徴を素性として抽出し、教師あり学習によって分類する手法を提案した。提案手法では、まず既存の日本語評価極性辞書をシードデータとした辞書の自動拡張を実施した。拡張には日本語 Wikipedia データを利用し、Word2Vec による分散表現と SVM を組み合わせることによって汎用的な拡張辞書を作成した。その後、議論コーパスの各固有表現に対し、語彙 N-gram、極性 N-gram、係り受け極性を定義し、RandomForest による学習・分類を実行した。提案手法の語彙 N-gram、極性 N-gram は有効であることが確認された。一方、係り受け極性及び辞書の拡張方法には課題が残る結果となった。

今後の展望として、辞書拡張時や素性作成時における不要な語句の判定及び除去、極性辞書以外によるポジティブ、ネガティブの判断を裏付ける表現の模索、また不均衡コーパスを補うためにポジティブ表現やネガティブ表現を他のデータから抽出することなどが挙げられる。

## 謝辞

本研究は、JST、CREST の支援を受けたものである。

## 参考文献

- [1] 伊藤孝行, 奥村命, 伊藤孝紀, 秀島栄三. 多人数ワークショップのための意見集約支援システム collagree の試作と評価実験:~議論プロセスの弱い構造化による意見集約支援~. 日本経営工学会論文誌, Vol. 66, No. 2, pp. 83-108, 2015.
- [2] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pp. 79-86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
- [3] Soroush Vosoughi, Helen Zhou, and Deb Roy. Enhanced twitter sentiment classification using contextual information. *CoRR*, Vol. abs/1605.05195, , 2016.
- [4] Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. Neural sentiment classification with user and product attention. In *EMNLP*, 2016.
- [5] Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. Attention-based lstm network for cross-lingual sentiment classification. pp. 247-256, 01 2016.
- [6] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pp. 347-354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [7] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pp. 226-234, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
- [8] Roy Bar-Haim, Lilach Edelstein, Charles Jochim, and Noam Slonim. Improving claim stance classification with lexical knowledge expansion and context utilization, 2017.
- [9] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 3, pp. 203-222, 2005.
- [10] 東山昌彦. 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会第 14 回年次大会論文集, 2008, pp. 584-587, 2008.