E-013

文法的な表現を手がかりとした宿泊施設レビュー文の意見分類
Classifying reputations of accommodations using grammatical expressions

大塚達也 † 立間淳司 ‡ 青野雅樹 ‡
Tatsuya Otsuka Atsushi Tatsuma Masaki Aono

1. はじめに

近年，インターネットやスマートフォンが普及し，個人が気軽に情報を発信できるようになった．発信される情報の一つとして，商品やサービスに対するレビュー文がある．商品やサービスを提供する企業は，レビュー文に書かれている苦情や要求を参考にして，商品やサービスを改良することができる．しかし，大量のレビュー文がある場合は，すべてを読むことは困難である．この問題に対して，レビュー文の意見を分類し，商品やサービスの改良を支援することを考えた．

2. 関連研究

レビュー文に関する研究には様々なものがある．Pang ら [1] は機械学習を用いて評価文章をポジティブとネガティブに分類している．小林ら [2] はレビュー文を〈対象，属性，評価〉の 3 つ組で定式化し，属性表現と評価表現を収集する手法を提案している．浅野ら [3] は正解ラベルとは別の談話素性ラベルや，前後の文を考慮してレビュー文の意見分類を行っている．

3. 提案手法

まず，与えられたレビュー文に対して形態素解析を行う．活用している単語は原型に変換する．得られた単語を用いて Bag-of-Words（BoW）や以下に述べる素性を計算し，学習器によって分類を行う．

3.1. 評価表現素性

評価表現が文中に現れた回数を素性にする．評価表現には，小林ら [2] の評価極性辞書を用いる．

3.2. 単語極性素性

「楽天データ公開」において公開されている楽天トラベルのレビューを用いて単語極性辞書を作成する．まず，楽天トラベルのレビューを Word2Vec [4] に入力し単語ベクトルを構築する．そして，構築した単語ベクトルとレビューに含まれる単語 w を用いて極性値 s を計算する．

$$s(w) = \mathrm{sim}\left(w, \frac{\alpha_1 \text{良い} + \alpha_2 \text{広い} + \alpha_3 \text{近い}}{\alpha_1 + \alpha_2 + \alpha_3}\right) - \mathrm{sim}\left(w, \frac{\beta_1 \text{悪い} + \beta_2 \text{狭い} + \beta_3 \text{遠い}}{\beta_1 + \beta_2 + \beta_3}\right)$$

ここで，sim() は単語ベクトルのコサイン類似度である．また，α_i，β_i は各単語の頻度である．

単語極性辞書を使用して，各文の単語に対しポジティブとネガティブの感情極性を付与する．文中に含まれるポジティブ単語とネガティブ単語をそれぞれ数え，以下の 4 つの素性を算出する．

1. ポジティブ単語数：p
2. ポジティブ単語数：n
3. ポジティブ単語数とネガティブ単語数の均衡：

$$balance_1 = \begin{cases} 0 & p = n = 0 \text{ のとき} \\ \exp\left(-\frac{(p-n)^2}{p+n}\right) & \text{それ以外のとき} \end{cases}$$

4. ポジティブ単語数とネガティブ単語数の調和平均：

$$balance_2 = \begin{cases} 0 & p = n = 0 \text{ のとき} \\ \frac{2pn}{p+n} & \text{それ以外のとき} \end{cases}$$

3.3. 文法表現素性

特定の文法的表現が現れた回数を素性とする．使用した文法項目は日本語能力試験出題基準の 3 級，4 級から計 11 個選択した．

例として，助動詞の「と」を使う条件表現は，以下の 4 つのパターンのいずれかが文中に現れたときに素性として用いる．

- 基本形動詞+接続助詞「と」
- 基本形形容詞+接続助詞「と」
- 未然形動詞+助動詞「ない」+接続助詞「と」
- 未然形形容詞+助動詞「ない」+接続助詞「と」

また，「ようになる」「ようにする」を使う変化表現は，以下の 2 つのパターンのいずれかが文中にあらわれたときに素性として用いる．

- 基本形動詞+名詞「よう」+助動詞語幹「に」+動詞「なる」
- 基本形形容詞+名詞「よう」+助動詞語幹「に」+動詞「する」

その他にも，表 1 に示す文法表現を素性として用いた．

4. 実験・考察

4.1. 実験データ

本研究では，「楽天データ公開」において公開されている筑波大学文単位評価極性タグ付きコーパスを用いた．このデータは，楽天トラベルの 1,000 件の施設レビューに含まれている 4,309 文に 2 人の作業者が，(褒め，苦情，ニュートラル，要求，評価なし) の 5 値の評価極性ラベルをつけたものである．実験では，与えられたレビュー文のラベルを正しく推定することを目的とする．

†豊橋技術科学大学 大学院工学研究科 情報・知能工学専攻
‡豊橋技術科学大学 情報・知能工学系

表 1: 評価表現辞書の例

素性	文法表現	単語
1	条件表現	と
2		ば
3		たら
4		なら
5	変化表現	なる，する
6		ようになる，ようにする
7	願望表現	ほしい
8		たい
9	逆説表現	ても
10		のに
11		けれども，けれど，けど，が

4.2. 実験方法

実験データは 4,309 文から無作為に選んだ 75%を学習データ，残りの 25%を評価用データとした．分類器は SVM を，カーネルは RBF カーネルを用いた．形態素解析器は MeCab を用いた．SVM のパラメータは，学習データ内で 5 分割交差検定を行いグリッドサーチで最良のものを選択した．

比較手法として，浅野ら [3] の手法を用いた．浅野らの手法においては，談話素性を SVM 分類器で自動推定した結果を比較対象とした．ベースライン素性として，頻度 3 以下の単語を除去した BoW，単語数，文の長さなどを使用した．BoW の単語の重み付けは，バイナリ値と TF-IDF 値を使用した．

4.3. 実験結果

4.3.1. 単語極性辞書

単語極性素性の計算において作成した単語極性辞書の単語極性値 s の上位 10 件と下位 10 件を表 2 に示す．ポジティブな単語として「便利」,「きれい」などの単語が取得できた．また，ネガティブな単語として「臭う」などの宿泊施設に関する単語が取得できた．

4.3.2. 意見分類の結果

意見分類の結果を表 3 に示す．BoW の重み付けに TF-IDF 値を用いたものは，すべての素性を組み合わせることでベースライン素性より 2.70%の向上が見られ，比較手法よりも優れた正解率となった．重み付けにバイナリ値を用いたものは，ベースライン素性より 2.23%の向上が見られた．

5. まとめ

本稿では，単語極性素性と文法表現素性を提案し，評価を行った．BoW とすべての素性を組み合わせた結果，従来手法よりも正解率において優れていることがわかった．今後の課題は，係り受け解析を用いて文法表現の検出ミスを少なくすることと，文章表現だけでは検出できないモダリティを解析して正解率を向上させることである．

表 2: 単語極性辞書

単語	品詞	極性値 s
良い-形容詞-自立	0.5046	
よい-形容詞-自立	0.4768	
好い-形容詞-自立	0.3636	
素晴らしい-形容詞-自立	0.3414	
いい-形容詞-自立	0.3248	
気持ち良い-形容詞-自立	0.3157	
すばらしい-形容詞-自立	0.3148	
心地よい-形容詞-自立	0.3068	
ここちよい-形容詞-自立	0.3038	
抜群-名詞-一般	0.3029	
⋮	⋮	⋮
かび臭い-形容詞-自立	-0.397	
淋しい-形容詞-自立	-0.3976	
匂う-動詞-自立	-0.4014	
さみしい-形容詞-自立	-0.4035	
斜め-名詞-形容動詞語幹	-0.4062	
におう-動詞-自立	-0.4312	
臭う-動詞-自立	-0.4476	
手狭-名詞-形容動詞語幹	-0.4606	
狭い-形容詞-自立	-0.4747	
せまい-形容詞-自立	-0.4838	

表 3: 実験結果（正解率（%））

手法	Binary	TF-IDF
比較手法	76.54	
BoW	75.12	74.91
BoW+評価表現	75.54	75.51
BoW+評価表現+単語極性	76.68	77.02
BoW+評価表現+単語極性+文法表現	77.35	77.61

参考文献

[1] 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.

[2] 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.

[3] 浅野翔太, 乾孝司, 山本幹雄. 談話役割に基づくクラス制約規則を利用したレビュー文の意見分類. 言語処理学会第 20 回年次大会 発表論文集, pp. 880–883, 2014.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, Vol. abs/1301.3781, , 2013.