レビューに対する分散表現を用いた評判分析

中嶌 大貴　　斎藤 博昭
慶應義塾大学 理工学部 情報工学科
{nakajima, hxs}@nak.ics.keio.ac.jp

1 はじめに

ウェブが普及し，様々なサービスや商品に関する評判が大量にブログやSNSに日々投稿されている．マーケティングをする上で製作者や販売者はこれらの情報を分析し利益につなげたいが，インターネット上に氾濫している大量の文書を人の手でチェックしていくのは非常に困難である．そのため，文書の内容を自動で判断する評判分析の需要は高く，これまでにも多くの研究が行われてきた．

既存の手法としてはBag of Wordsを素性として，人の手で単語の評価を記録した評価極性辞書を用いて評判分析を行う手法が多かったが，辞書作成に人的コストがかかることや比較的新しい語に対応できないという欠点があった．CBOWおよびSkip-gramという手法を用いて単語の特徴抽出を行う分散表現という手法がMikolovら [3] によって提案され，これらの欠点を解決することができるため近年活発に研究が行われている．

これまで，日本語の文書に対する分散表現を用いた評判分析については新聞記事のようなある程度書式が統一されているデータに関する実験は多く行われているが，Q & Aサイトや掲示板のようにユーザが自由に記述できるメディア (CGM：Consumer Generated Media) に対して分散表現を用いた評判分析を適用したものは少ない．そこで本論文ではCGMであるホットペッパービューティーの口コミデータに対して分散表現を適用し，2値分類の評判分析を行い効果を確認する．今回は分散表現に対する比較対象として，文書ベクトルの伝統的な表現方法である1-of-K表現に，次元に対応する単語のTFIDF値を重み付けしたものを用いる．

2 関連研究

本節では，本研究を行う背景となった研究について述べる．

評判分析における評価極性分類タスクにおいて,「不満」という否定的な意味を持つ単語に「減る」という単語が加わり「不満が減る」という肯定的な意味の文節ができてしまうといったような単語間の相互作用の問題がある．Nakagawaら [4] はこのような評価極性が反転してしまう事例に対応するために，評価表現の依存構造木について個々の部分依存構造木に対する評価極性を隠れ変数で表現することで，評価極性分類タスクの分類精度を上げることに成功した．しかし，Nakagawaらの手法はタスクを行うにあたって12種の組成テンプレートを作る必要があるため複雑なモデルであるといえる．

評価極性分類タスクにおいて，否定的なことを大量に述べた後に「けど」などの逆説から肯定的なことを述べることで文章の極性が肯定になるというように，文脈によって評価極性が反転してしまう問題がある．Ikedaら [2] はこのような評価極性が反転してしまう事例に対応するために，単語の周辺を考慮した学習手法として各単語について極性が反転しているか否かを学習する単語極性反転モデルを提案している．

これらの研究はいずれもモデルが複雑であるが，本論文が提案するモデルは非常に単純であり実装が容易である．

また，評価極性分類を用いてユーザの傾向を識別するといったタスクが存在する．TASS2013ワークショップは2013年のスペインの選挙について著名人158人がスペイン語で書いたツイートを用いて著名人の政治傾向を識別する精度を競うが，Plaら [5] はツイート内の文章を政治傾向毎に分ける処理を行った後，各文章に対してポジティブ，ネガティブ，中性の極性分類を行うことで最も良い精度を出すことに成功した．上記の研究 [4][2][5] は全て評価極性辞書を用いたものであり，評価極性辞書に載っていない比較的新しい語に

は対応できない．

評価極性辞書を用いずにマイクロブログの評価極性分類を行った研究としては dos Santos ら [1] や Tang ら [6] の研究が挙げられる．dos Santos らは畳み込みネットワークを文字単位で用いて文ベクトルを作成したのに対し，Tang らは単語の分散表現自体を極性として表現したモデルを使用したが，どちらの研究も本論文とは文ベクトルの作成方法が異なる．また張ら [7] は評価極性辞書を用いずに新聞記事に対して評価極性分類を行った．張らの研究は分散表現による文ベクトルの作成方法は本論文の提案手法と同じであるが，本研究は書式が統一されていないデータに対して分散表現を用いて評判分析を行ったという点で異なる．

3 提案手法

3.1 提案手法の概要

本提案手法の概要を図 1 に示す．提案手法ではレビューデータを入力とし，入力されたレビューデータから分散表現で特徴を抽出する．抽出された特徴からレビューデータの 2 値分類を行い，分類結果を本提案手法の出力とする．

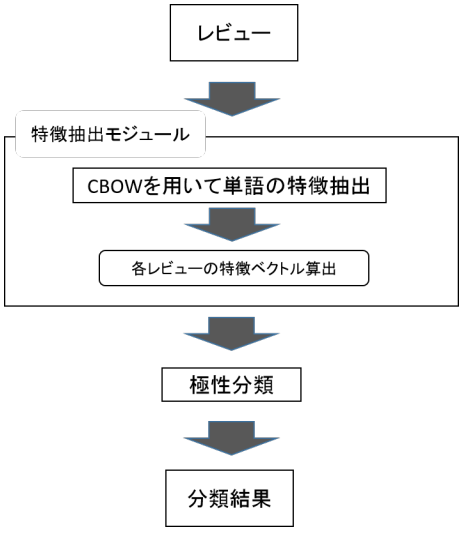

図 1: 提案システムの概要

3.2 単語の特徴抽出

これは分散表現を用いて特徴抽出を行う場合のみ必要となる処理である．Continuous Bag-of-Words Model というニューラルネットワーク言語モデルの手法を使用する．図 2 に CBOW の概要を示す．CBOW モデルは入力層，射影層，出力層の 3 層からなり，単語列 W における w_t の周辺単語 $w_{t-k}, ..., w_{t-1}, w_{t-1}, ..., w_{t+k}$ を入力として w_t を予測するニューラルネットワークである．CBOW の入出力は単語の 1-of-V 表現であり，射影層では入力された各単語の 1-of-V 表現を d 次元のベクトルに射影して和を計算する．CBOW では誤差逆伝搬法を用い，射影層への重みを学習する．学習終了後は入力単語の 1-of-V 表現と射影層への重みから，入力単語の特徴ベクトルを得ることができる．

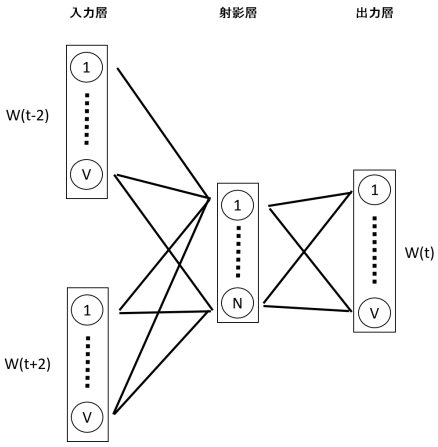

図 2: CBOW の概要

3.3 レビューの特徴ベクトル算出

処理が異なるため，レビューの特徴抽出に分散表現を使う場合と 1-of-K 表現を使う場合に分けて説明する．

分散表現を用いる場合

レビュー集合 S に含まれる 1 件のレビュー S_k の特徴ベクトルは S_k に含まれる単語 $(w_i)^{(k)}$ の分散ベクトルを用いて，以下の (1) 式によって算出される．ただし，N_k は S_k に含まれる単語の数である．

$$S_k = \frac{1}{N_k} \sum_{j=1}^{N_k} (w_j)^{(k)} \quad (1)$$

これは S_k に含まれる単語の分散ベクトルの平均である．

1-of-K 表現を用いる場合

レビュー集合 S に含まれる 1 件のレビュー S_k の特徴ベクトルは，レビュー集合 S に含まれる単語の種類数を n とすると以下の (2) 式によって算

出される．ただし，x_i には対応する次元の単語が S_k に含まれていなければ 0 が，S_k 含まれていれば対応する単語 $(w_i)^{(k)}$ の TFIDF 値が入る．

$$S_k = [x_1, ..., x_i, ..., x_n] \tag{2}$$

(2) 式より算出したベクトルは次元数が非常に大きくなってしまうため，実験の際は主成分分析で次元数を削減する．

算出されたレビューの特徴ベクトルについては，z-Score を用いて正規化を行う．ベクトル集合 X に含まれるベクトル x の z-Score は以下の (3) 式によって求めることができる．ただし $\overline{X}$ はベクトル集合 X の平均，v はベクトル集合 X の不偏標準偏差である．

$$Z = \frac{x - \overline{X}}{v} \tag{3}$$

3.4 極性分類

分類方法を 1 つにしてしまうと特徴抽出方法との相性が悪いことも考えられるため，今回は SVM およびロジスティック回帰の 2 種類の識別方法を用いることとした．

4 実験

4.1 実験設定

実験データ

実験用のレビューデータは ADVANCED TECHNOLOGY LAB がリクルートオープンデータとして公開しているヘアサロン検索サイト，ホットペッパービューティーの口コミデータを用いる。この口コミデータにはニックネーム・性別・世代・レビュー内容・総合評価などのデータが含まれているが、本実験ではレビュー内容と総合評価のみを抜き出して使用した。評価は 1 から 5 までの 5 段階評価であり，5 が一番良い評価である．評価値に対応したレビューを表 1 の件数用意した．

表 1: 使用したデータセットの件数

評価値	1	2	3	4	5
レビュー数	5000	5000	0	5000	5000

計 20000 件のレビューデータに対して，評価 1 と 2 のデータについては False，評価 4 と 5 のデータについては True のタグ付けを行う．タグは分類結果に対する正解データとして用いる．本実験では 2 値分類を行うため，中間値である評価 3 のデータは使用しない．ベクトル化を行うにあたって入力を単語に区切る必要があるため，レビューデータは MeCab[1] を用いて形態素解析を行うとともに，不要な単語であると考えられるため句読点を取り除いた．3.3 節で述べた 2 種類の手法を用いて，形態素解析を行ったレビューデータのベクトル化を行った．ベクトル化する際には Word2Vec[2] と scikit-learn[3] を使用した．1-of-K 表現については主成分分析を用いて次元削減することで各レビューデータに対して 200 次元からなる 2 種類の特徴ベクトルを得た．

分類手法と調整

SVM のパラメータについては，4.1 節で用意したデータセットとは別に，チューニング用のデータセットを表 2 の件数用意し，グリッドサーチを用いて最適なパラメータを探索した．ロジスティック回帰についても，表 2 で用意したチューニング用のデータセットを用いてグリッドサーチで最適な学習率を探索した．

表 2: チューニング用データセットの件数

評価値	1	2	3	4	5
レビュー数	1000	1000	0	1000	1000

4.2 実験結果

実験は K-分割交差検証で行い，K の値は 15 とした．実験の結果を分類方法毎に分けて表 3，表 4 に示す．正解率は実際の分類結果とタグが一致した場合を正解とし，正解と不正解を足した数で割ったものである．

表 3: SVM を用いた場合の分類結果

ベクトル	正解率	F1
分散表現	0.953	0.950
TFIDF	0.766	0.758

[1] http://taku910.github.io/mecab/
[2] https://radimrehurek.com/gensim/index.html
[3] http://scikit-learn.org/stable/

表 4: ロジスティック回帰を用いた場合の分類結果

ベクトル	正解率	F1
分散表現	0.953	0.950
TFIDF	0.757	0.750

5 考察

表3，表4より，分類方法にSVMおよびロジスティック回帰のどちらを用いても分散表現を用いて特徴抽出した方が1-of-K表現にTFIDF値を重み付けしたものを用いるよりも約0.2ポイント精度が高いという結果になった．これは分類手法に依らず分散表現を用いた特徴抽出が有用であるということを示している．

分類方法を比較すると，1-of-K表現で特徴抽出した場合SVMを用いた方がロジスティック回帰よりも約0.01ポイント精度が高く，分散表現で特徴抽出した場合でも精度が変わらないという結果となった．今回はSVMのパラメータ探索にグリッドサーチを使用したが，グリッドサーチでは重要なパラメータを取りこぼすことがあるため，ランダムサンプリングを用いてパラメータ探索を行うことで更に精度が上がる可能性がある．

6 おわりに

本稿では分散表現を用いて単語の特徴抽出を行い，抽出した単語ベクトルを元に文ベクトルを作成することで評判分析を行う方法を提案した．ホットペッパービューティーの口コミデータに対して適用し，従来の1-of-K表現を用いた特徴抽出法に比べ精度が高く有用であることを示した．

今回評価に使用した元文書の文字数は100-300文字程度と比較的短いものだった．より文字数が多い文書の場合の検証や，肯定及び否定だけでなく中性の判別もできるように機能を拡張すること，文字の順序を考慮してより精度を上げる特徴抽出方法の設計が課題として挙げられる．今後は様々な文書データで有用性を検証し，トピック分類と組み合わせることで文書の自動分類を行うシステムの構築など応用先を見つけていきたい．

謝辞

本研究では，株式会社リクルートテクノロジーズが国立情報学研究所の協力により研究目的で提供している「リクルートデータセット」を利用させていただいた．

参考文献

[1] Cicero dos Santos and Maira Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 69–78, 2014.

[2] Daisuke Ikeda, Hiroya Takamura, Lev-Arie Ratinov, and Manabu Okumura. Learning to shift the polarity of words for sentiment classification. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, pp. 296–303, 2008.

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

[4] Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 786–794, June 2010.

[5] Ferran Pla and Lluís-F. Hurtado. Political tendency identification in twitter using sentiment analysis techniques. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 183–192, 2014.

[6] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1555–1565, 2014.

[7] 培楠張, 守小町. 単語分散表現を用いた多層 denoising auto-encoder による評価極性分類. Technical report, 首都大学東京システムデザイン研究科, 首都大学東京システムデザイン研究科, 2015.