7P-04

分野別感情極性辞書の作成及び評価

森田 晋也　白井 靖人

静岡大学 大学院 総合科学技術研究科情報学専攻

1. はじめに

情報インフラの発達にともない,大量かつ多様なテキストが蓄積されるようになり,それらを活用しようという試みが各分野にて行われている.その中でもテキストから物事に対する評価や感情を判定する評価情報分析に関する研究が数多くなされている. 評価情報分析には単語の感情極性を利用する方法がある.感情極性とはある単語が一般的にいい意味を持つか,悪い意味を持つかを表した二値属性である.単語の感情極性を判定する方法として,語彙ネットワークを利用した方法[1]や,単語の共起情報を利用した方法[2]など,いくつか提案されている.中でも単語の分散表現と機械学習を用いた方法[3]は,構築コストが低く応用の範囲が広い.

しかし,いずれの方法でも単語の多義性については考慮されていない.そこで単語の多義性は分野を限定することで低減できると考え,本論文では評価極性分析の精度を上げるために,特定の分野に着目し,その分野での使われ方を考慮した単語感情極性辞書の作成とそれを用いて評価情報分析を行った.

なお対象とするのは日本語テキストである.

2. 基盤技術など

本論文では以下の技術やパラメータを用いる.

2.1.1 Word2vec[4]

分散表現を学習するモデルには CBOW,Skipgram, 及びGloveを使用する.windows∈{5,10},dimension∈{100,200,400,800,1000}で試行する.

2.1.2 SVM

分野ごとの単語の分散表現を素性とし,感情極性を予測するように分類機に学習させる.その際,学習データには既存の辞書を使用する.SVM には線形カーネルと RBF カーネルを使用し,それぞれパラメータは C∈{1,10,100,1000},RBF カーネルは gamma∈{0.001,0.0005,0.0001}で試行する.

2.1.3 意見(評価表現)抽出ツール

評価情報分析には独立行政法人情報通信研究機構 旧知識処理グループ情報信頼性プロジェクトによって開発された意見(評価表現)抽出ツール[1]を使用する. 本ツールは 1 行につき 1 文が書かれたテキストファイルを入力として,その文に評価情報が存在すると判定した場合,評価表現・評価タイプ・評価極性・評価保持者を出力する.

2.2 分野別感情極性辞書の作成

機械学習の結果から得られた単語の感情極性を用いて,以下の方法に従い二種類の分野別感情極性辞書を作成する.

2.2.1 既存の辞書の更新

既存の辞書に掲載されている単語の感情極性と機械学習によって得られた感情極性が異なっていた場合,既存の感情極性を得られた感情極性で更新する. 見出し語の追加はしない.

2.2.2 既存の辞書の更新と見出し語の追加

2.2.1 のように更新を行った後,単語の追加を行う. その際,重要度の高い単語は評価極性分析を行う際に有用であると考え,TFIDF 値がその分野の TFIDF 値の平均値より大きい単語を追加する.

3. 結果

分野別の日本語テキストには楽天株式会社が提供している楽天データの楽天市場の商品レビュー[2]を使用した.機能語は除き,活用形は終止形に戻した.また出現頻度が小さい単語は取り除いた.

SVM の学習データには 2.1.3 に付属する辞書データ 35,000 語(pos:10,000 語 neg 25,000 語)を使用した.

評価は評価極性付きテキストである KNB コーパス[5]の評価極性を正解とし,意見(評価表現)抽出ツールを使用しそのテキストから得られた評価極性を比較する.

意見抽出ツールに既存の辞書を適用した場合(Baseline)と生成した分野別感情極性辞書を適用した場合の正答率を比較する.

---

[1]意見（評価表現）抽出ツール
https://alaginrc.nict.go.jp/opinion/extractopinion-1.1/index.html

[2]楽天データ
https://rit.rakuten.co.jp/data_release/

一文に対し評価極性が複数付与されている場合,多い方の評価極性をその文の評価極性とする.同数であった場合は,末尾の評価極性をその文の評価極性とする.

3.1 更新した既存の辞書を用いた場合

2.2.1 にて述べた方法で生成した辞書を適用し評価を行った．辞書に登録されている単語数は Baseline と同じ 36,981 語である.表 3-1 は各分野のベースラインの正答率と適用した辞書の中で最も良い正答率になった結果を掲載した.

表 3-1　3.1 各分野正解率

分野		正解	不正解	正解率
機械	Baseline	250	329	0.43
	Challenge	252	327	0.44
グルメ	Baseline	148	227	0.39
	Challenge	151	224	0.40
スポーツ	Baseline	94	145	0.39
	Challenge	96	143	0.40

3.2 更新と追加を行った辞書を用いた場合

2.2.2 にて述べた方法で生成した辞書を適用し評価を行った.各分野追加された単語数は, 機械分野で 1,882 単語,グルメ分野で 1,962 単語,スポーツ分野で 1,892 単語である.表 3-2 は各分野のベースラインの正答率と作成した辞書の中で最も良い正答率になった結果を掲載した.

表 3-2　3.2 各分野正解率

分野		正解	不正解	正解率
機械	Baseline	250	329	0.43
	Challenge	324	255	0.56
グルメ	Baseline	148	227	0.39
	Challenge	207	168	0.55
スポーツ	Baseline	94	145	0.39
	Challenge	110	129	0.46

4.　考察

既存の辞書を更新する方法では正解率に大きな違いはみられなかった. 一方,既存の辞書に更新と追加を行う方法では正解率の向上が見られた.

また今回,様々なモデルやパラメータを使用したが,2.2.2 の方法の場合,機械分野は分散表現のモデルに Glove を用いたものの正解率が高く,スポーツは CBOW を用いたものの正解率が高い傾向があった.しかし共通して特定のモデル,パラメータが適している等の傾向は見られなかった.

5.　終わりに

本論文では特定の分野での使われ方を考慮した感情極性辞書を作成することで,その分野での評価情報分析の精度が向上すると考え,2 種類の方針の辞書を作りそれぞれを用いて評価を行い,比較した.

その結果,辞書を更新するだけでは正解率を向上させることはできず,見出し語を追加することで正解率を向上できることが分かった.

さらなる検証のため,感情極性を得ることができた単語全てを更新・見出し語追加の対象とした辞書を用いて同様の実験を行なった. 各分野追加された単語数は，機械分野で 6,370 単語,グルメ分野で 8,693 単語,スポーツ分野で 6,905 単語である表 5-1 はその結果である.

表 5-1 追実験 各分野正解率

分野		正解	不正解	正解率
機械	Baseline	250	329	0.43
	Challenge	332	247	0.57
グルメ	Baseline	148	227	0.39
	Challenge	216	159	0.58
スポーツ	Baseline	94	145	0.39
	Challenge	112	127	0.47

3.2 と比べ大きな変化は見られなかった.この結果から,辞書への見出し語の追加による正解率の向上にはある程度効果があると考えられるが,3.2 の時点である程度頭打ちであったと思われる.

参考文献

[1] 高村大也, 乾孝司, 奥村学. "スピンモデルによる単語の感情極性抽出" ,情報処理学会論文誌ジャーナル,Vol.47 No.02 , pp. 627—637,2006.

[2] Truney,P.D. and Littman M.L. “Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus.” Tech. rep., Technical Report NRC Technical Report ERB-1094 , Institute for Information Technology , National Research Council Canada, 2002.

[3] 佐藤貴俊,高村大也,奥村学 “分散表現を用いた単語の感情極性抽出”,情報法処理学会研究報告,Vol.2016-NL-228.No12,2006

[4] Tomas Mikolov , Kai Chen ,Greg Corrand, and Jeffery Dean. “Efficient Estimation of Word Representations in Vector Space.”, CoRR, abs/1301.3781, 2013.

[5] 橋本力, 黒橋禎夫, 河原大輔, 新里圭司, 永田昌明. “構文・照応・評判情報つきブログコーパスの構築.” 自然言語処理,Volume 18, Number 2, pp.175-201,2011.