

Twitter のリプライを用いた感情辞書の自動生成

大前 遼太郎[†] 杉本 徹[‡]芝浦工業大学大学院 理工学研究科[†] 芝浦工業大学 工学部[‡]

1. 研究背景と目的

自然言語処理において、入力したテキストを解析し、感情を推定する技術は商品レビューの分析などで使用されている。テキストの感情を推定するアプローチの1つとして感情辞書を用いる方法がある。感情辞書に関する研究として佐藤ら[1]の研究がある。この研究では単語をポジティブとネガティブに分類し単語の極性を求めている。しかしポジティブとネガティブへの分類だけでは感情の分類としては十分とは言えない。その理由として単語が持つ感情は必ずしも1つではないからである。「山」という単語は登山などから楽しいという感情を連想する人も一方で、遭難から怖いという感情を連想する人もいる。単語をどれか1つの感情に分類するよりも、それぞれの感情に対してどれくらいの程度を持っているのかを定義する方が人の感情として自然である。複数感情の辞書の構築を行う研究として武内らの研究[2]がある。この研究では、ニューラルモデルを用いてコーパスから単語の感情ベクトルを取得する。得られた感情ベクトルと人間による感情評価とのコサイン類似度は0.695であった。

本研究では、Twitter のリプライには元ツイートの内容に対する感情が表現されやすいという仮説のもと、ツイートとリプライを一種の感情タグ付きコーパスとみなす。このコーパスを用いてツイートに出現する単語が表す感情の推定を行う。本手法で推定した単語の感情と人間による判断との比較を行い、提案手法の検証と考察を行う。

2. 研究内容

2.1. 感情の定義

感情の分類はいくつかの方法が提案されている。その中でも中村 [3] は言語表現の観点から、感情を「喜、怒、哀、怖、恥、好、厭、昂、安、驚」の10感情に分類している。本研究ではこの分類手法を用いる。

Automatic generation of an emotion lexicon using
Twitter replies

[†]Ryotaro Omae, [‡]Toru Sugimoto

[†]Graduate School of Engineering and Science, Shibaura
Institute of Technology

[‡]College of Engineering, Shibaura Institute of Technology

2.2. ツイートの収集

感情表現辞典 [3] では、10感情を表す単語や熟語が収録されている。その中で各感情の意味をあらわす語が見出しに書かれているため、それらの語を抜き出した。例えば「喜」の感情では「めでたい、うれしい、満足、楽しい、快感、微笑」という語を抜き出した。これら抜き出した語が含まれるツイートを各感情に対して収集を行った。収集を行ったツイートは全感情合わせて約500万個であり、そのうち本研究で使用するリプライは約50万個である。

2.3. 感情ベクトルの作成

収集したツイートのうちリプライであるものに対して、その元ツイートを取得し、MeCab[4]を用いて単語ごとに分割する。感情ベクトルを作成するために自己相互情報量 PMI を以下の式で計算する。

$$PMI(e, w) = \log \left(\frac{T * \text{count}(e, w) + 1}{\text{count}(e) * \text{count}(w)} \right)$$

PMI(e, w) = 単語 w の感情 e の値
T = 取得したすべての元ツイートに含まれる単語の総数
count(e, w) = 感情 e の語を含むリプライの元ツイートにおける単語 w の出現回数
count(e) = 感情 e の語を含むリプライの元ツイートの単語の総数
count(w) = 取得したすべての元ツイートにおける単語 w の出現回数

単語 w に対して10感情それぞれに対して求めた PMI(e, w) の値を並べたベクトルを単語 w の感情ベクトルとする。

3. 評価実験

3.1. 実験の内容

作成した感情ベクトルの妥当性を評価するために、感情ベクトルと人間による判断を比較する実験を行う。評価に用いる単語は、システムが出力した感情ベクトルの少なくとも1つの感情の値が1.0以上の単語15個とすべての感情が1.0未満の単語15個をランダムに選択した。表1に具体例を示す。

大学生13人に、各単語から10感情それぞれが

どの程度想起されるかを 0~4 の 5 段階で回答してもらいアンケートを行った。感情ごとに全回答を足し合わせたものを並べたベクトルを評価ベクトルと呼ぶ。

表 1. 評価に用いた単語の例

1.0 以上の単語例	1.0 以下の単語例
芸人、悲報、平気	レース、雑魚、初期

3.2. 結果と考察

3.2.1 単語分散表現を用いた感情ベクトル

ベースラインとして単語分散表現を用いた感情ベクトルの作成を行った。単語分散表現は Word2Vec[5] (日本語 Wikipedia 本文データ / Skip-gram /100 次元) で学習したものである。各単語に対し、10 感情それぞれに関して、ツイート収集の時に用いた感情の意味を表す語とのコサイン類似度を求め、コサイン類似度の最大値をその単語におけるその感情の値とする。これを 10 感情分並べたものをその単語の感情ベクトルとする。

3.2.2 ベクトルの組み合わせ

リプライを用いて求めた感情ベクトルと単語分散表現を用いて求めた感情ベクトルを組み合わせることで新たに感情ベクトルを作成する。この時ベクトルを各ベクトルの成分の最大値が 1、最小値が 0 となるようにベクトルごとに正規化した。感情ごとに 2 つのベクトルにおける値の最大値 (または最小値、平均値) を求めたものを 10 個並べてできるベクトルを最大値ベクトル、最小値ベクトル、平均値ベクトルと呼ぶ。

3.2.3 結果と考察

評価ベクトルと新たに作成した感情ベクトルのコサイン類似度を求めた結果を表 2 に示す。

表 2. コサイン類似度

感情ベクトルの作成手法	コサイン類似度
リプライを用いる手法	0.682
単語分散表現を用いる手法	0.696
最大値ベクトル	0.736
最小値ベクトル	0.695
平均値ベクトル	0.739

リプライを用いる手法と単語分散表現を用いる手法を組み合わせる最大値ベクトルや平均値ベクトルで高いコサイン類似度が得られた。

評価ベクトルと感情ベクトルにおける最も値の大きい感情の一致率を表 3 に示す。

表 3. 最も値が大きい感情の一致率

リプライを用いる手法	26.6%
単語分散表現を用いる手法	13.3%

次に感情ごとの傾向を調べるために、評価ベクトル、リプライおよび単語分散表現を用いて

作成した感情ベクトルにおける、全単語にわたっての各感情の順位の平均を表 4 に示す。

表 4. 各感情の順位平均

感情	評価ベクトル	リプライを用いる手法	分散表現を用いる手法
喜	3.7	6.1	4.7
怒	6.3	6.9	5.7
哀	6.3	4.5	6.1
怖	5.3	5.0	8.2
恥	6.9	4.8	5.6
好	4.0	6.4	5.7
厭	5.7	4.4	4.2
昂	5.0	5.2	5.5
安	5.3	5.9	6.2
驚	6.0	6.4	4.6

今回評価に使った単語では人は喜の感情が 3.7 と少し想起しやすかった。一方、リプライを用いる手法では喜の感情は 6.1 と順位が低い傾向が見られた。また、分散表現を用いる手法では怖の感情が 8.2 と順位が低かった。2 つの手法で傾向が異なるため、平均値をとることによって互いの苦手な部分を補うことでコサイン類似度が高くなったと考えられる。

4. まとめと今後の展望

本研究では Twitter のリプライを用いて感情ベクトルの作成を行う手法を提案した。単語分散表現を用いた手法と組み合わせることにより人間の判断とのコサイン類似度が 0.739 と単独の手法の場合より高くなった。一方、リプライを用いる手法単独では単語分散表現を用いる手法に及ばなかった。その理由の一つとして考えられるのは、使用するコーパスの大きさの違いである。今後は収集に使う語の変更や収集数を増やすなどをしてさらなる改善を目指したい。

参考文献

- [1] 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”, 情報処理学会研究報告, Vol. 2016 NL228, No. 12, 2016
- [2] 武内達哉, 萩原将文, “単語の持つ感情推定法の提案と単語感情辞書の構築”, 日本感性工学会論文誌, Vol. 18, No. 4, pp. 273-278, 2019
- [3] 中村明, “感情表現辞典”, 東京堂出版, 1993
- [4] <https://taku910.github.io/mecab/>
- [5] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, “Distributed representations of words and phrases and their compositionality”, NIPS, 2013