

# Twitter を利用した感情辞書の自動生成

大前 遼太郎<sup>†</sup> 杉本 徹<sup>‡</sup>

芝浦工業大学大学院 理工学研究科<sup>†</sup> 芝浦工業大学 工学部<sup>‡</sup>

## 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いる一方で、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、特定のキーワードが含まれるTwitterのツイートを感情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を様々な手法で推定を行う。様々な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

<sup>†</sup>Ryotaro Omae, <sup>‡</sup>Toru Sugimoto

<sup>†</sup>Graduate School of Engineering and Science, Shibaura Institute of Technology

<sup>‡</sup>College of Engineering, Shibaura Institute of Technology

### 2.2. ツイートの収集

感情付きコーパスを生成するために、顔文字が含まれたツイート、絵文字が含まれたツイート、感情表現辞典[3]から選んだ単語が含まれた単語を感情毎にそれぞれ集めた。そのツイート群で予備実験を行い、ツイート群の中で感情が正しい割合の多いツイート群に対して前処理を行いコーパスとした。コーパスのサイズは全感情合わせて約51万ツイートとなった。

### 2.3. 感情ベクトルの作成

コーパスから単語の感情を推定する手法を以下に説明する。

#### 2.3.1. 単語と感情の共起情報を用いた方法

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。

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

PMI(e, w) = 単語 w の感情 e の値

T = コーパスに含まれる単語の総数

count(e, w) = 感情 e の語を含むツイートにおける単語 w の出現回数

count(e) = 感情 e の語を含むツイートの総数

count(w) = コーパスにおける単語 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。

#### 2.3.2. 単語分散表現を用いた方法

各単語に対し、10 感情それぞれに関して、感情表現辞典[3]から選んだ単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。この時使用した単語分散表現は Word2Vec[5] (日本語 Wikipedia 本文データ / Skip-gram / 200 次元) で学習したものである。

#### 2.3.3. ニューラルネットワークを用いた方法

入力層は語彙数が次元数であるベクトルであり、これは語彙のリストのインデックスに対応している。入力層と中間層の間の重み w1 は、語

彙数を  $h$ 、感情の分類数を  $e$  とすると、 $(h \times e)$  の行列で表せる。出力は入力層と重み  $w_1$  を掛け合わせた和である。この値は、入力した単語それぞれが持つ感情の値の和に等しい。出力層は 1 ツイートを入力した際に推定される感情の値を表す。この出力値と出力の教師データである感情について、交差エントロピーを最小化するように学習を行う。学習を行ったモデルに対し、入力層と中間層の間の重みの値を取得する。これにより得られる値を感情ベクトルとする。

### 2.3.4. アテンション付きニューラルネットワークを用いた方法

上記のモデルの中間層の後に教師データとの内積を取り、入力ごとの重みを計算するアテンション層を追加する。出力は中間層と重みを掛け合わせた和である。2.3.3 節と同様に入力層と中間層の間の重みの値を取得する。これにより得られる値を感情ベクトルとする。

### 2.3.5. 複数の方法の組み合わせ

2.3.1~2.3.4 節で述べた方法で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。この時手法毎にベクトルの成分の最大値が 1、最小値が 0 となるように正規化した。その後様々な方法で新たなベクトルを作成した。予備実験により最も高い結果が得られた 2.3.1 節と 2.3.3 節で作成したベクトルの各成分の平均を取ったベクトルを使用する。

## 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 2.3.5 節の予備実験とは異なる 30 単語とする。評価単語の例を表 1 に示す。

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平均し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

|                    |
|--------------------|
| 芸人、悲報、平気 レース、雑魚、初期 |
|--------------------|

### 3.2. 結果と考察

評価ベクトルと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 |
|-------------|---------|
| 手法 2.3.1    | 0.809   |
| 手法 2.3.2    | 0.812   |
| 手法 2.3.3    | 0.807   |
| 手法 2.3.4    | 0.797   |

|          |       |
|----------|-------|
| 手法 2.3.5 | 0.809 |
|----------|-------|

手法 2.3.2 が最も高いコサイン類似度が得られた。

次に、各単語  $W_i$  に対して評価ベクトルで最大の値をもつ感情が作成した感情ベクトルの全成分の中で何番目に大きい感情であるかを調べ、平均逆順位 (MRR) を求めた結果を表 3 に示す。

表 3. 各手法の MRR

| 感情ベクトルの作成方法 | MRR   |
|-------------|-------|
| 手法 2.3.1    | 0.338 |
| 手法 2.3.2    | 0.364 |
| 手法 2.3.3    | 0.304 |
| 手法 2.3.4    | 0.322 |
| 手法 2.3.5    | 0.343 |

MRR も手法 2.3.2 が最も大きくなった。どの手法も MRR が 0.3 以上であり、評価ベクトルにおける最大の感情が平均的に 3 位以内に入っていることが分かった。

各感情の精度を求めるため、手法 2.3.2 の感情毎のベクトル差の絶対値の平均を求めたところ最小が「好」で値が 0.178、最大が「哀」で値が 0.300 となった。

どの評価指標でも手法 2.3.2 がよい結果が得られた。そこで 2.3 節で使用したツイートで改めて Word2Vec[5] (Skip-gram /200 次元) で学習を行い類似度を求めた。結果は 0.790 となり最も類似度が低くなった。ここから考えられる理由としてはコーパスサイズである。Wikipedia のコーパスが約 3 GB であるのに対しツイートコーパスは約 190MB と大幅に少ない。そのため十分な学習ができず類似度が低くなったと考えられる。

## 4. まとめと今後の展望

Twitter を用いて感情ベクトルの作成を行う手法を提案した。Wikipedia の単語分散表現を用いた手法が人間の判断とのコサイン類似度が 0.812 と最も高くなった。しかし同じコーパスならば単語と感情の共起情報を用いた方法が最も類似度が高くなった。今後はコーパスの拡大や収集語のさらなる検討により精度を高めていきたい。

## 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, 18(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 et al., “Distributed representations of words and phrases and their compositionality”, NIPS, 2013