2R-01

構成要素に着目した顔文字の意味分析

源田　翼†　　横井　健†　　山本　和英‡
東京都立産業技術高等専門学校 ものづくり工学科 電子情報工学コース†
長岡技術科学大学 電気電子情報工学専攻‡

1　はじめに

マイクロブログサービス等の文字数の少ないテキストにおいては、感情や行動を簡潔に表現するために顔文字がしばしば用いられる。それらの顔文字の種類は現時点においても豊富に存在し、なおかつ新たな顔文字が追加され続けている。そこで、あらゆる顔文字に対して、その顔文字が表現する感情や行動を柔軟に分析するため、本研究では顔文字を構成する要素に着目する。

顔文字は、目や口などを表現する文字を構成要素として、それらの組み合わせによって感情や行動を表現している。この事実に基づいて、本研究では各構成要素の意味を分析することにより、顔文字全体が表現する感情や行動を分析する手法について検討する。

2　関連研究

顔文字が表現する感情や行動の分析については、これまでにもさまざまな手法が提案されている。その内、本研究では、近年、自然言語処理の分野で注目を集めている Word2Vec[1]というものを用いて感情分類を行っている黒崎らの研究[2]に注目した。

Word2Vec とは、「同じ文脈で出現する単語は同じ意味を持つ」という仮説を基に、単語をベクトル化して表現する手法である。「書店・へ・行く」のように、単語単位で分かち書きした文書を Word2Vec へ入力すると、入力文書内の各単語に対して、その意味を表現したベクトルが出力される。

黒崎らの研究では、まず顔文字を単語の一つとみなして、それらを含んだ大規模な文書を Word2Vec へ入力する。その後、出力されたベクトル集合から顔文字の表現している感情を分析している。しかし、この手法では、ベクトル集合を作成する際に Word2Vec へ入力した文書内に存在する顔文字に対してしか分析が行えない。そこで、顔文字の種類が増加した場合でも分析する手法が求められる。

Semantic Analysis of Emoticon Based on the Structural Elements
†Tsubasa GENTA, Takeru YOKOI
‡Kazuhide YAMAMOTO
†Tokyo Metropolitan College of Industrial Technology
‡Nagaoka University of Technology

3　提案手法

顔文字の構成要素の中には各感情と強い共起関係にあるものが存在する[3]。このことから、顔文字の構成要素はそれぞれが意味を持っていると考え、各構成要素の意味を分析することにより、顔文字全体が表現する感情や行動を分析する手法を提案する。

3.1　提案手法の流れ

提案手法の流れを以下に示す。

(1) 顔文字が含まれた文書中から顔文字を抽出し、文書中に存在しない固定長のハッシュに置換する。
(2) MeCab を用いて文書を単語と顔文字のハッシュごとに分かち書きする。
(3) 文書中の顔文字のハッシュを、その顔文字の構成要素ごとに分かち書きしたものに置換する。
(4) 文書中の単語と顔文字の構成要素のベクトル表現を、Word2Vec を用いて作成する。
(5) 構成要素のベクトル表現をもとに顔文字のベクトルを表現する。
(6) 顔文字のベクトル表現と、後述する各感情を代表する単語のベクトル表現とのコサイン類似度を計算する。
(7) 顔文字とのベクトル表現同士の類似度が最も高い単語が代表している感情を、その顔文字が表現している感情として分類する。

3.2　分類する感情の種類

本実験で分類を行う感情は Plutchik の感情モデル[4]を参考に、「喜び」と「悲しみ」、「怒り」と「恐れ」の 4 種類とした。

また、各感情を代表する単語は、「喜び」、「悲しみ」、「怒り」、「恐れ」のそれぞれについて、「うれしい」、「かなしい」、「怖い」、「おこる」とした。

3.3　顔文字のベクトル表現

構成要素のベクトル表現をもとに顔文字のベクトルを表現する手法を以下に示す。

3.3.1　構成要素のベクトルの総和

顔文字eのベクトル表現$\boldsymbol{v}_e$は式(1)で求める(提案手法 1)。

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.

$$\boldsymbol{v}_e = \sum_{c \in C} n_c \cdot \boldsymbol{v}_c \qquad (1)$$

式(1)におけるCは顔文字eを構成している文字の集合、n_cはその顔文字中での文字cの出現回数、$\boldsymbol{v}_c$は文字cのベクトル表現を表す。

3.3.2 各構成要素の出現頻度による重み付き総和

各構成要素の顔文字中での逆出現頻度ief_c(Inverse Emoticon Frequency)を式(2)で定義する。

$$ief_c = \log \frac{|E|}{|\{e|e \ni c\}|} \qquad (2)$$

式(2)における$|E|$はWord2Vecへ入力した文書中の顔文字の種類数、$|\{e|e \ni c\}|$は文字cを含んでいる顔文字の種類数を表す。

これを用いて、顔文字eのベクトル表現$\boldsymbol{v}'_e$は式(3)で求める(提案手法 2)。

$$\boldsymbol{v}'_e = \sum_{c \in C} ief_c \cdot n_c \cdot \boldsymbol{v}_c \qquad (3)$$

4 実験

提案手法の性能を評価するために感情分類の実験を行った。

4.1 実験方法

文書には、Twitterのパブリックタイムラインより無作為に取得した 15,751,004 件の日本語ツイートから URL やハッシュタグ、その他 Twitter 特有の表現を除去したものを使用した。

また、顔文字の抽出は人手によって行い、352 種類の顔文字を使用した。

4.2 評価手法

抽出した顔文字を人手によって各感情に分類した。352 種類の顔文字の内、139 種類が「喜び」を、81 種類が「悲しみ」を、82 種類が「怒り」を、50 種類が「恐れ」を表現していた。その後、提案手法を用いてそれらの顔文字の感情分類を行い、適合率と再現率、そして適合率と再現率の調和平均である F 値を求めた。また、比較のために、人の手で分類した中で最も顔文字の種類が多かった感情「喜び」に全ての顔文字を分類した場合(比較手法 1)と、ランダムに分類した場合(比較手法 2)の各値を求めた。

4.3 実験結果

各手法による分類の適合率を表 1 に、再現率を表 2 に、F 値を表 3 にそれぞれ示す。

両提案手法とも比較手法 1 に比べ、F 値の平均は高くなった。また比較手法 2 に比べ、「喜び」や「悲しみ」に対して高い F 値が得られた。

しかし、「怒り」と「恐れ」の F 値が低かったため、その平均は比較手法 2 とほぼ等しくなった。

4.4 考察

結果のように「怒り」と「恐れ」の F 値が低くなった原因は、これらの顔文字を構成している文字に統一性があまりなく、他の感情を表現する顔文字でも多く用いられるためであると考えられる。また、同じ文字が異なる物体を表すことがあるのも原因の一つだと考えられる。例えば、「悲しみ」を表現する顔文字内で涙を表すためによく用いられる「;」は、「恐れ」を表現する顔文字内では汗を表すために用いられることがある。分類結果でも、誤って「悲しみ」に分類された顔文字の内、「恐れ」を表現している顔文字は汗を表すために「;」を使用しているものが多かった。そういった場合には、構成要素一文字単位ではなく、いくつかの集まりを一つの単位として扱うなどの必要があると考えられる。

表1 各手法の適合率

手法	喜び	悲しみ	怒り	恐れ	平均
提案手法1	**0.80**	0.27	0.14	**0.27**	**0.37**
提案手法2	0.74	**0.28**	0.15	0.23	0.35
比較手法1	0.39	0.00	0.00	0.00	0.10
比較手法2	0.43	0.27	**0.24**	0.13	0.27

表2 各手法の再現率

手法	喜び	悲しみ	怒り	恐れ	平均
提案手法1	0.26	**0.93**	0.01	0.12	0.33
提案手法2	0.32	0.88	0.02	0.12	**0.34**
比較手法1	**1.00**	0.00	0.00	0.00	0.25
比較手法2	0.24	0.32	**0.26**	**0.24**	0.26

表3 各手法のF値

手法	喜び	悲しみ	怒り	恐れ	平均
提案手法1	0.39	0.42	0.02	0.17	0.25
提案手法2	0.45	**0.43**	0.04	0.16	**0.27**
比較手法1	**0.57**	0.00	0.00	0.00	0.14
比較手法2	0.31	0.29	**0.25**	**0.17**	0.25

5 おわりに

本研究では顔文字の構成要素に着目し、各構成要素が表現している意味から顔文字の感情や行動を分析することを試みた。

その結果、「喜び」と「悲しみ」に対してはその有用性を確認できた。しかし、「怒り」と「恐れ」に対しては、分類が正しく行えなかったため、その改善手法について検討を行う必要がある。

参考文献

[1] Mikolov, Tomas. Sutskever, Ilya. Chen, Kai. Corrado, Greg. and Dean, Jeffrey. Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26. Curran Associates, Inc., 2013, p. 3111-3119.

[2] 黒崎優太, 高木友博. 言語処理学会第 21 回年次大会. 言語処理学会, 2015, p. 441-444.

[3] 山本湧輝,熊本忠彦, 灘本明代. Twitter 特有表現を考慮したツイートの多次元感情抽出手法.情報処理学会関西支部支部大会講演論文集.情報処理学会関西支部, 2014.

[4] Plutchik, Robert. The nature of emotions. American Scientist, 2001, vol. 89, no. 4, p. 344-350.

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.