3H3-OS-17a-4

Twitterにおける顔文字を用いた感情分析の検討

Study of Sentiment Analysis using Emoticons on Twitter

風間 一洋 *1 水木 栄 *2 榊 剛史 *2*3
Kazuhiro Kazama Sakae Mizuki Takeshi Sakaki

*1和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University

*2株式会社ホットリンク
Hotto Link Inc.

*3東京大学工学系研究科
School of Engineering, The University of Tokyo

Emoticons, which look like human faces, are symbol sequences to express user's sentiments of their text messages. Because users create new emoticons or modify existing emoticons easily, too many variants are produced continously. There are some researches of sentiment analysis using emoticons, it is difficult for them to treat such problems. In this paper, we discuss a sentiment analysis framework using emoticons. To create a large emoticon sentiment dictionary, we extract many different kind of emoticons automatically by using information such as Unicode character properties and Unicode blocks. Furthermore, we extract two types of clusters from various emoticons: one is synonyms of a emoticon and another is synonymies of a emoticon.

1. はじめに

インターネットの普及に伴い，電子メールや電子掲示板，ブログ，SNS などの主にテキストを用いたメッセージ交換がさかんに行われるようになった．ただし，相手の表情を見ることができないことから感情を正確を伝えることが難しく，誤解が生じないように著者の意図を顔文字（emoticon, smiley）で補足することが一般的に行われてきた．逆に，顔文字に着目すれば，テキストに込めた感情をより正確に推定できる．

顔文字を用いて感情分析を行うためには，あらかじめ顔文字極性辞書を構築しておく必要がある．さらに，顔文字極性辞書を日本語形態素解析の辞書と統合すれば，日本語形態素解析と顔文字抽出を同時に行うことができる [榊 16]．しかし，ユーザが顔文字を容易に創造・変更できるために，種類が膨大なことに加えて，絶え間なく変種が生み出されている．さらに OS ベンダの Unicode サポートの充実により，文字の動的合成を駆使して多彩に表現する特殊顔文字も急速に普及しつつある．つまり，実用に耐える顔文字極性辞書を構築するためには，顔文字の多様性と時間的変化に追従できる必要がある．

本稿では，顔文字極性辞書を自動的に構築するために必要となる技術について議論する．まず大規模なコーパスから，特殊顔文字を含む従来よりも多彩な顔文字を自動的に抽出できる手法について述べる．次に，抽出された膨大な種類の顔文字を，同義顔文字と類義顔文字というグループにまとめて，顔文字の基本形と感情語または感情分類を持つ顔文字極性辞書を構築するための表層的類似性と意味的類似性を判定する手法について述べる．

本研究では，これらの技術に加えて，表層的特徴と意味的特徴を組み合わせた顔文字の感情推定を実現することで，大規模・多様な顔文字に対する極性辞書の自動生成を可能にし，顔文字の多様性と時間的変化に追従できる実用的な感情分析フレームワークの構築を目指す．

連絡先: 風間 一洋 (kazama@ingrid.org)
和歌山大学システム工学部
〒 640–8510 和歌山県和歌山市栄谷 930

2. 顔文字抽出

2.1 顔文字の自動抽出法

大規模な顔文字辞書や顔文字極性辞書を構築するために，様々な顔文字の自動抽出法が提案されている．

まず，顔文字を構成する部品（目，口，輪郭など）の配置や対称性を仮定して抽出する方法が存在する．Ptaszynski らの CAO システムでは，目–口–目の三つ組に着目して顔文字を抽出した [Ptaszynski 10]．Bedrick らは，顔文字を構成する Unicode 文字のグリフの類似性と鏡像関係に注目して，PCFG モデルを用いて顔文字候補を抽出した [Bedrick 12]．三好らは，顔文字内を文字 n-gram の出現頻度に基づいて感情分類するための分類済の顔文字コーパスとして，顔文字以外にテキストや修飾記号なども多く含まれる顔文字図書館 *1 のデータから，正規表現を用いて半角または全角の丸括弧で囲まれている文字列だけを顔文字として抽出した [三好 13]．

また，与えられた顔文字集合の出現パターンを学習して，顔文字を抽出する手法が存在する．Tanaka らは，茶筌で日本語形態素解析した後で文字単位で付与された形態素の種類・形態素中の位置，文字種などのタグ情報を素性とし，SVM を用いた分類器である Yamcha を用いてチャンキングして顔文字を抽出した [Tanaka 05]．渡邉らは，学習用データに人手で文字ごとに FACE-B（顔文字の先頭文字），FACE-I（顔文字の 2 文字目以降），TEXT（顔文字以外の文字），EOS（文末）の 4 種類のラベルを付与し，文字と文字種の情報を素性として CRF で系列ラベリングをおこなって FACE-B に FACE-I が 0 文字以上連続している文字列を顔文字として抽出した [渡邉 13]．

ただし，先行研究では，顔文字を構成する部品に関するさまざまな仮定や学習に使用した顔文字集合に抽出結果が制約される上に，Unicode の動的文字合成を用いた場合に対応できない．そこで，形状や配置には踏み込まずに仮定や制約を大幅に緩和し，顔文字を文章とは異なる文脈で存在する記号や日本語以外の文字を多く含む文字列とみなし，ルールベースの処理で顔文字候補を抽出する [風間 13]．このために，Unicode Standard で定義されている Unicode 文字プロパティと Unicode ブロッ

*1 http://www.kaomoji.com/kao/text/

表 1: 記号類と判定する Unicode 文字プロパティ

Lm	Letter, Modifier
Pc	Punctuation, Connector
Pd	Punctuation, Dash
Pe	Punctuation, Close
Pf	Punctuation, Final quote
Pi	Punctuation, Initial quote
Po	Punctuation, Other
Ps	Punctuation, Open
Sc	Symbol, Currency
Sk	Symbol, Modifier
Sm	Symbol, Math
So	Symbol, Other

表 2: 日本語と判定する Unicode ブロック

ブロック名	コード範囲		
Basic Latin	U+0000	〜	U+007F
Hiragana	U+3041	〜	U+309F
Katakana	U+30A0	〜	U+30FF
CJK Unified Ideographs	U+4E00	〜	U+9FFF
Fullwidth ASCII Variants	U+FF01	〜	U+FF60
Halfwidth Katakana Variants	U+FF61	〜	U+FF9F

クを使用する．また，顔文字の左右に手や腕を表す文字がくる場合に関しては，与えられた顔文字リストから特別扱いするべき部分文字列を自動抽出して対処する．最後に顔文字以外の記号列を除去することで，多彩な顔文字の抽出を実現する．

2.2 顔文字抽出アルゴリズム

実際には，以下の 1 から 6 の処理を抽出されたテキストの終端に達するまで繰り返し，顔文字である領域を決定する．

1. **顔文字探索**: テキストの前方から顔文字主要文字を探す．終端に到達したら終了．
2. **領域拡張**: 見つかった場合は，顔文字主要文字の間に最大 G 個の任意の文字を許容しながら，領域を前後に拡大する．
3. **領域縮小**: 先端・終端の括弧・句読点または顔文字主要文字以外の削除を繰り返し，領域を縮小する．
4. **領域補完**: 小規模な顔文字リストから事前に抽出した部分文字列を用いて欠落部分を判定・補完する．
5. **顔文字判定**: 領域が L 文字以上の場合は，それが顔文字かどうかを判定する．
6. 判定済み領域の次の文字に移動し，1 に戻る．

1，2，3 で用いる顔文字主要文字は，顔文字で主に用いられている文字である．ツイートの文章は，ある特定の言語で記述されるが，顔文字はそれと明らかに区別がつくような記号やそれ以外の言語の文字が多用される傾向がある．そこで，Unicode Standard の文字の機能を定義する Unicode 文字プロパティと文字の種類を定義する Unicode ブロックを用いて，Unicode 文字プロパティの一般カテゴリ [*2] が表 1 に示す値を持つ記号類か，表 2 に示す日本語の文字以外の文字を，顔文字主要文字とする．

*2 http://www.unicode.org/reports/tr44/#General_Category_Values

表 3: 文字と判定する Unicode 文字プロパティ

Lc	Letter, Cased
Ll	Letter, Lowercase
Lm	Letter, Modifier
Lo	Letters, Other
Lt	Letter, Titlecase
Lu	Letter, Uppercase
Nd	Number, Decimal Digit
Nl	Number, Letter

2 で最大 G 個の任意の文字を許容する理由は，例えば「(T_T)」の「_」や「(; ;)」の空白のような，記号以外の文字の混在をある程度許容するためである．ただし，G の値を増やすと，通常の単語が記号類と一緒に誤抽出される可能性が高くなる．

3 では，顔文字は文末に使われる傾向があるために「(^^)『」や「。(^^)」のような前後の括弧や句読点と隣接することが多いので，領域の前後でそれぞれ一部括弧類の向きが異なる 38 文字を削除する．Unicode 文字プロパティを使わなかった理由は，その分類よりもさらに記号が限定されるからである．

4 は，iOS でサポートされている 137 文字の顔文字リストを，顔文字の前後に対して，3 の処理を施した後で，顔文字と判定されなかった部分を隣接する 1 文字と共に抽出して，拡張用の部分文字列として使用する．例えば「m(_ _)m」からは左側の「m(」と右側の「)m」が抽出される．ただし，「！ (◎_◎;)」に関しては，文末の！ を顔文字の一部として許容すると，その有無で顔文字の種類が増えるので，「(◎_◎;)」として扱った．この結果抽出された部分文字列は，顔文字の左側が 12 種類，右側が 18 種類である．

5 の顔文字の判定は，以下の経験則に基づく．

1. 同じ文字の繰り返しではない（例，*****）．
2. 両側が「と」，『と』，”と”で囲まれていない．
3. 丸括弧の中に数字や漢字がある文字列ではない（例，(1)，(株)）．
4. Unocode 文字プロパティで表 3 と判定される文字の比率が半分より小さい．これはハングルなどの海外の文字で書かれたツイートに対する対策である．

なお，膨大なデータを処理すれば顔文字と記号類の隣接による誤抽出の頻度は正しく抽出される場合よりも低くなり，出現頻度が高い顔文字は比較的正しく抽出される．

2.3 提案法の特徴

顔文字の両側を補完するために顔文字リストを使用しても，人手の選別やタグ付けは不要である．

また，提案法では複数の顔を用いた場合にはまとめて抽出する（例，「o(^-^o)(o^-^)o」）が，周囲の修飾部分も同時に抽出できる（例，「。・゜・(ノД‘)・゜・。」），

提案法で顔文字の領域を決定した後に，それ以外のテキスト部分を日本語形態素解析することで，解析ミスを減らすと共に，顔文字が直前のテキストの感情を補完すると仮定することで，関連する感情語の抽出や感情の推定ができる．

3. 顔文字を用いた感情分析に向けて

顔文字を手掛かり情報として用いることで，感情分析を行う手法は今までも行われてきた [Read 05, Go 09]．これらの多くは，顔文字を文書に対する感情のラベルや手掛かりとみなし

表 4: 抽出した顔文字サンプルとその出現日時

顔文字	初出月
＼(^o^)／◎	2007年 3月
＼(^o^)／⁉	2007年 4月
♬＼(^o^)／	2007年 6月
＼(^o^)／♫	2007年11月
＼(^o^)／☀	2008年 3月
`・ω・´☆	2007年 2月
(∪´・ω・`∪)	2007年 2月
(∪*´ω`)	2009年 1月
ヾ(´ω`)*	2009年 2月
(*´ω`)つ【コーヒー】	2012年 3月
('Θ')	2007年11月
('Θ')ノ	2008年 7月
∈('Θ')∋	2010年 9月
∈('Θ')∋☆	2010年11月
Λ('Θ')Λ”	2011年 8月

て，教師有り学習や Distant Supervision を適用するアプローチ [Hu 13, Tang 14] である．本稿では，顔文字の多様性と時間的変化に追従できることを目的としているが，既存のアプローチはあくまで所与の顔文字を扱うのみであり，未知の顔文字の出現に対処することは困難であると推測される．

以下では，顔文字自体の変化にも対処可能なアプローチを検討する．

3.1 表層的類似性と意味的類似性

すでに多彩な顔文字を抽出できる方法について述べたが，顔文字には頻繁に新しいパターンが作りだされる特徴がある．例えば，表 4 に類似した顔文字とその初出日を示す *3 ．このように同系統の顔文字において多様なパターンが存在し，新パターンの出現頻度も一般的な感情語の場合より非常に多い．

顔文字の感情分析の先行研究においては，1 つの顔文字に 1 つの感情が割り当てられる (例,「(^ ^)/」はポジティブ,「(>_<)」はネガティブ) と仮定することが多い [Read 05, Go 09]．実際，殆どの顔文字は単一の意味で用いられる事が多く，そのようなアプローチは有効である．なお，極性のようにポジティブ・ネガティブ・感情なしの 3 分類以外にも，Plutchik の感情の輪などに代表される有限個のグループに分類して扱われることが多いので，提案した顔文字抽出法で抽出された多彩な顔文字をグループ化して扱うことが必要となる．

本稿では，グループ化における類似性の判定基準として，意味的類似性と表層的類似性の 2 種類を考える．

意味的類似性とは，見かけは異なる顔文字の意味的な類似性のことである．意味的に類似した顔文字を類義顔文字と呼び，いずれも見かけは違っても同じ意味で用いられていると推測される．類義顔文字のグループは，感情分析で使用される感情分類に相当する．

表層的類似性とは，ほぼ同じ意味で用いられる顔文字の見かけ (表層) の類似性のことである．表層的に類似した顔文字を同義顔文字と呼び，いずれも同じ顔文字から派生したものと推測される．同義顔文字のグループは，テキスト分析で使用される表記ゆれに相当する．例えば，奥村は顔文字の原型抽出とそれに用いるルールについて提案し，人手でアノテーションを行っている [奥村 16] が，同義顔文字を適切に抽出できれば，このような作業を自動化できるだけでなく，顔文字パターン数の削減や新たな顔文字グループの誕生の発見などに使用できる．

*3 ホットリンク社のクチコミ係長のブログデータで，各顔文字の出現頻度の 180 日移動平均値が 30 件/月を超えた日を初出日とした

意味的類似性を用いれば，表層的な類似性の有無にかかわらず，顔文字の類似性を扱うことができる．このため，顔文字を扱う上では，意味的類似性のみを考慮すれば問題がないように見える．しかし，意味的類似性を扱うためには，対象となる顔文字そのものが出現するコーパスが大量に必要となるアプローチが殆どである．つまりコーパスに出現しない未知の顔文字を扱うことが困難である．他方，表層的類似性を扱う場合は，対象となる顔文字の部分 (文字 n-gram など) が出現するコーパスが用意できればよい．そのため未知の顔文字にも対応できる可能性が高い．これより，表層的類似性と意味的類似性は相補的に活用することで，より頑健に顔文字の類似性を扱えると考えられる．

以下で，この 2 種類の類似性の判定法について述べる．

3.2 表層的類似性の判定

表層的類似性は，顔文字の文字 n-gram の類似性に基づいて判定する．具体的には，全顔文字集合 U に対して，その中から選択した m 個の顔文字 $(t_1, t_2, \ldots, t_m)$ が与えられたときに，顔文字の特徴を文字 n-gram によって表現した顔文字－n-gram 行列を作成し，さらに NMF（Non-negative Matrix Factorization）を用いて低次元空間に射影することで次元を削減し，与えられた顔文字との類似度をその空間上で計算することで同義顔文字かどうかを判定する．詳細な手順を次に示す．

1. 顔文字集合 U に含まれる全ての顔文字 $e_i (1 \leq i \leq |U|)$ から n-gram$(n = 1 \sim 2)$ を抽出して，n-gram 特徴の集合 G を作成する．
2. 各顔文字 e_i を，それを構成する n-gram の $|G|$ 次元のベクトル $(g_{i,1}, g_{i,2}, \ldots, g_{i,|G|})$ で表現して，$|U|$ 行 $|G|$ 列の顔文字－ n-gram 行列 X を構成する．
3. 行列 X を，NMF を用いて $|U|$ 行 K 列の行列 T と K 行 $|G|$ 列の行列 V に分解する．

$$X = T \times V \tag{1}$$

ここで $K = m$ とする．この結果，T の j 行目が顔文字 $t_j (0 \leq j \leq m)$ に対応する K 次元の表層的特徴ベクトル v_{t_j} となる．

4. 顔文字 t_j との L2 距離が閾値 th を越えた顔文字 e_i の集合を，顔文字 t_j の同義顔文字 E_j として抽出する

$$Sim(e_i, t_j) = cos(v_{e_i}, v_{t_j}) \tag{2}$$

$$E_j \rightarrow Synonym(if\ Sim(e_i, t_j) > th) \tag{3}$$

上記手順について，一定手順 *4 によって抽出した 20 個の顔文字に対して抽出した同義顔文字を表 5 に示す．定量的な評価を行っていないが，同じ顔文字から派生したと想定される顔文字を抽出することができた．これらの定量的な評価は今後の課題であるが，表層的な類似性については本アプローチで抽出可能であることが示唆された．またソーシャルメディア上での出現頻度や出現時期など一定の基準を用いて同義顔文字の基本形および活用形を定めることも可能である．

3.3 意味的類似性の判定

意味的類似性は，「意味的な類似性」に分布仮説を適用し，単語分散表現の一つである word2vec[Mikolov 13] を用いることで得られる「使用される文脈の類似性」に基づいて判定する．具体的には，全顔文字集合 U に対して，その中から選択した

*4 Twitter での出現頻度によって顔文字を順位付けした後，rank mod 50 =1 となる上位 20 語を抽出した．

表 5: 抽出された同義顔文字リスト

グループID	1	2	3	4	5	6	7	8
顔文字リスト	＼(^o^)／	m(_ _)m	(−−;)	(^。^)	(´ω`*)	ϵ(´Θ`)з	(*･ω･)	ヽ(≧▽≦)/
	／(^o^)＼	m(_ _m)	⁉(−−;)	(^^)。	(*´ω`)	ϵ(´Θ`)з”	(･ω･*)	ヽ(〃▽〃)/
	＼(^o^)／･	%m(_ _)m	⁇(−−;)	。(^^)	(*`ω´)	ϵ(´Θ`)з“	(･ω･)*	ヽ(≧∇≦)/
	＼(^o^)／✧	m(_ _)m※	(−−;)	(。^∪^)	(`ω´*)	ϵ(´Θ`)з<	](･ω･*)	ヽ(≧▽≦)/♪
	℃＼(^o^)／	m(_ _)m‖	~(−−;)	(^。^)❤	(´ω`)*	(´Θ`)	ϕ(･ω･*)	♪ヽ(≧▽≦)/

表 6: 抽出された類義顔文字リスト

グループID	1	2	3	4	5
人手ラベル	困った系	喜び系	悲しみ系	その他系	挨拶系
顔文字リスト	(´･_･`)	＼(^o^)／	(>_<)	(`･ω･´)	(^^)
	(´Д`)	(*^^*)	(T_T)	(･∀･)	(^-^)
	(´･ω･`)	(^o^)	(;_;)	(`･ω･´)	(^-^)/
	(^_^;)	(((o(*▽*)o)))	ヽ(;▽;)ノ	(･ω･)	(^_^)
	(´Д`)	(*´ω`*)	(´;ω;`)	(*∀*)	m(_ _)m

m 個の顔文字 $(t_1, t_2, \ldots, t_m)$ が与えられたときに，各顔文字の特徴を word2vec を用いて分散表現で表した後で，与えられた顔文字との類似度をその空間上で計算することで類義顔文字を抽出する．詳細な手順を次に示す．

1. Twitter データから，提案法で抽出した顔文字と，それ以外のテキストを日本語形態素解析して得られた単語から，word2vec を用いて単語分散表現を構築する．
2. 顔文字集合 U に含まれる顔文字 $e_i (1 \leq i \leq |U|)$ に対して，単語分散表現の意味的特徴ベクトル v'_{e_i} を抽出する，
3. 顔文字 t_j とのコサイン類似度が閾値 th を越えた顔文字 e_i の集合を，顔文字 t_j の類義顔文字 E_j として抽出する．

$$Sim(e_i, t_j) = cos(v'_{e_i}, v'_{t_j}) \tag{4}$$

$$E_j \rightarrow Synonym(if\ Sim(e_i, t_j) > th) \tag{5}$$

上記手順に基づいて，一定基準で抽出した顔文字 1000 語の類義顔文字クラスタを生成した．結果を表 6 に示す．定量的な評価は行っていないが，意味的に類似していると推測される顔文字が同じクラスタに集約されていることから，定性的には本手法がうまく機能していることが推測される．

4. おわりに

本稿では，顔文字を用いて感情分析を実現するためのフレームワークについて議論した．まず，大規模コーパスから多彩な顔文字を自動抽出する手法と，顔文字の基本形と感情語を対応づけた顔文字辞書を用いる方法について述べた．さらに，抽出された膨大な種類の顔文字を同義顔文字と類義顔文字というグループにまとめて扱うための表層的類似性と意味的類似性を判定するアプローチについて提案した．

今後は，表層的特徴と意味的特徴を組み合わせて，新たに出現した顔文字の表層的な特徴からその顔文字が持つ感情（意味的特徴）を推定するアプローチに取り組む予定である．

参考文献

[Bedrick 12] Bedrick, S., Beckley, R., Roark, B., and Sproat, R.: Robust kaomoji Detection in Twitter, in *Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)*, pp. 56–64 (2012)

[Go 09] Go, A., Bhayani, R., and Huang, L.: Twitter sentiment classification using distant supervision, *CS224N Project Report, Stanford*, Vol. 1, p. 12 (2009)

[Hu 13] Hu, X., Tang, J., Gao, H., and Liu, H.: Unsupervised Sentiment Analysis with Emotional Signals, in *Proceedings of the 22nd International Conference on World Wide Web*, pp. 607–618 (2013)

[風間 13] 風間 一洋, 榊 剛史, 鳥海 不二夫, 篠田 孝祐, 栗原 聡, 野田 五十樹：顔文字に着目したツイートの感情変化の分析, *WebDB Forum* (2013)

[Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, in *Proceedings of the 27 Annual Conference on Neural Information Processing Systems (NIPS2013)* (2013)

[三好 13] 三好 辰明, 太田 学：ツイートに出現する顔文字等の文字と記号に着目した感情分類, *DEIM Forum 2013 D9-2* (2013)

[奥村 16] 奥村 紀之：顔文字の原形抽出, 言語処理学会第 22 回年次大会 (NLP2016) (2016)

[Ptaszynski 10] Ptaszynski, M., Maciejewski, J., Dybala, P., Rzepka, R., and Araki, K.: CAO: A Fully Automatic Emoticon Analysis System Based on Theory of Kinesics, *IEEE Transactions on Affective Computing*, Vol. 1, No. 1, pp. 46–59 (2010)

[Read 05] Read, J.: Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification, in *Proceedings of the ACL Student Research Workshop(ACLstudent '05)*, pp. 43–48 (2005)

[榊 16] 榊 剛史, 水木 栄：ソーシャルメディア分析サービスにおける NLP に関する諸問題について, 言語処理学会第 22 回年次大会ワークショップ「論文に書かない（書けない）自然言語処理」(2016)

[Tanaka 05] Tanaka, Y., Takamura, H., and Okumura, M.: Extraction and Classification of Facemarks with Kernel Methods, in *Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI '05)*, pp. 28–34 (2005)

[Tang 14] Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B.: Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL2014)*, pp. 1555–1565 (2014)

[渡邉 13] 渡邉 謙一, 高橋 寛幸, 但馬 康宏, 菊井 玄一郎：系列ラベリングによる顔文字の自動抽出と顔文字辞書の構築, 言語処理学会第 19 回年次大会 (NLP2013), pp. 866–869 (2013)