

# 顔文字情報と文の評価表現の関連性についての一考察

村上 浩司      山田 薫      萩原 正人

楽天株式会社 楽天技術研究所

{koji.murakami, kaoru.yamada, masato.hagiwara}@mail.rakuten.co.jp

## 1 はじめに

これまで我々が手に出来る文字情報として、情報伝達を目的とした新聞記事などの書き言葉に加えて、人間同士のコミュニケーションを文字化した話し言葉もある。現在、多くの話し言葉から構成されるブログや Twitter などの個人から発信される情報 (CGM) が爆発的な規模で日々増加しており、含まれる表現は多岐に渡る。自然言語処理も同様にその解析対象を書き言葉から話し言葉へと広げることで、多様な言語表現解析、深い意味的な解析などが可能になってきた。CGM に含まれる情報は必ずしも言語表現だけではなく、それ以外の顔文字や感情文字、所謂ギャル文字などの記号列も多く含まれ、これらは言語情報そのものや文字だけでは表現しにくい書き手の意図、感情などの情報を示していると考えられる。

これまで、こうした情報を適切に処理し利用してきた研究はそれほど多くなかったが近年、顔文字や記号などの非言語情報、不自然言語といわれる文字列を言語処理の一部として解析対象とする動きが出てきた。本論文ではこうした不自然言語と呼ばれる文字列の中でも、複数の記号から構成される「顔文字」に着目した。極性辞書の構築とその利用を目的として、Twitter から抽出した顔文字を対象に顔文字の極性判定実験の結果から、顔文字情報を適切に捉えるための多クラス分類の必要性、事例分析や関連研究について報告する。

本論文は、2 節は顔文字を対象としたこれまでの研究について、3 節は顔文字辞書構築の基礎実験結果を述べる。4 節において顔文字情報を適切に扱うための方法についての議論を行い、5 節でまとめを述べる。

## 2 研究対象としての顔文字

顔文字とは、“(^\_^)” や “\ (^o^)/” のように表情やしぐさに似せて、記号や文字を組み合わせた列であり、一般的には文末に付与することでその文の感情極性や意味に対して付加的な情報を与える目的で用いられる [5]。日本で顔文字が初めて出現したのは 1986 年頃<sup>1</sup>とされ、その利用が一般的になってから大凡 10 年ほど経過すると思われる。個人的な意見を書き込むことのできる情

報媒体であるブログ、2ch、Twitter、掲示板などで頻繁に目にする。

心理学や認知科学など分野では 90 年代後半から顔文字に着目してきた。

- (1) 旅行から帰ったら銀行すっからかん (!\_+) [パラ言語情報の付与]
- (2) それって当たり! !(^~)! [強調]
- (3) あの本持ってきてくれない? [配慮]
  - a. 分かった
  - b. 分かった (^o^)

上で示すような文が現れる文字コミュニケーション中において、言語表現で表すことのできないプロソディ要素やパラ言語的要素などの微妙な感情表現を補う情報である [2, 3] として、強調 [11] や感情 [13, 12, 11]、配慮 [3]、解釈 [10] などの役割について研究が行われてきた。自然言語処理分野では、井上ら [1] が顔文字などの記号列が感情を伝達することを示し、記号を含む感情表現の分類、分析を行った。また田中らにより、文からの顔文字抽出および 6 種類の感情カテゴリへの高精度での分類が報告されている [14]。しかしながら言語処理分野においては、これまで言語情報に解析対象として焦点があり、感情情報分析であっても顔文字を対象にしたものは多くなく、むしろ不要な文字列として削除される傾向にあった。顔文字のほかにも、句点としての音符 (「」) や文末表現に付随する文字列 (例えば「w, ww」)、感情文字 (例えば「(汗)」なども不要な情報として扱われてきた。

顔文字は、ただの意味のない記号列とも捉えられるが、その一方、先に示した例のように、話者の感情や意図などを扱う際に非常に重要な情報を持つことが分かる。また、顔文字単体で示す極性情報なども重要であるが、顔文字とそれが出現する文の感情情報との関連により表される強調や配慮などの情報もあることから、文との関連性についても着目する必要がある。

## 3 顔文字辞書構築のための基礎実験

前節で述べたように、顔文字は書き手の感情や意図を認識するために非常に重要な情報となりえる。顔文字

<sup>1</sup><http://www.ja.wikipedia.org>

Annotator	Positive	Negative	Neutral	Total
A	333	217	323	873
B	250	234	389	873
Common	219	144	194	557

表 1: 顔文字に対する極性付与結果

が感情情報を持つことから、ここでは顔文字極性辞書を想定して、構築のための基礎実験を行う。言語表現を対象とした感情分析はこれまで、品詞や句に対する正 (Positive)、負 (Negative)、中性 (Neutral) の 3 クラス極性辞書の構築が行われてきた (例えば [6, 9] など)。顔文字に対しても同様の辞書を考える。

### 3.1 顔文字の収集

まず顔文字を収集する必要がある。顔文字は前述の通りブログや掲示板など、Web 上の情報の多くに存在している。我々は最近ユーザが多く、さまざまな顔文字が出現する Twitter に着目し、弊社サービスであるロボット“らくったー”<sup>2</sup>のフォロワーのツイート約 127 万件を解析対象とした。これらのツイートは 2009 年 7 月から 2010 年 9 月までの 1 年 2 ヶ月で収集されたものである。ユーザ数 4,315、1 ユーザ当たりの最大ツイート数 15,609、最小ツイート数 1、平均 296 ツイートである。

顔文字であるか否かは、正解顔文字集合が存在しないため単純に選別することができない。そこで、つぶやき中に存在する言語情報以外の記号列の頻度を計測する。高頻度である場合、それらが意図を持って使われていると考え、顔文字と認定することとする。まず全ツイートから平仮名、片仮名、漢字を含む日本語部分を削除して残った部分を顔文字候補とした。もし顔文字が日本語文字を含む場合はこの方法では獲得できないが、まずは取得容易な顔文字候補の抽出に主眼を置くこととする。このとき多くの種類の記号列が得られることが考えられるため、意味のないと考えられる 2 文字以下の記号列は候補の対象外にする。この処理により、4,420 の候補が抽出された。

顔文字候補はあくまで記号列からのみ構成されることから、全く偶然的に並んだ記号列も存在する。そこで顔文字が“顔”であるという事実に基づいて、顔文字らしさの低いものを人手によってフィルタリングする。顔文字らしさとして、以下のようなガイドラインを設け、これらの基準を満たさないものを顔文字ではないと判断した。

- 顔の構成要素 顔を構成する要素が存在する (例えば、目:(鼻|口):目の順と分かる記号列など)
- 顔の輪郭 括弧により顔らしさを確認できる、もしくは記号列だけで顔と認識可能

人手によるこの作業により約 1,500 種類の記号列が得られ、更に頻度が 4 以上であるもの 873 種類をの最終的な

(*_~*)	( ' ' )	\ ( ' O ' ) /
(~_*)	( ~ ~ )	\ ( ~ o ~ ) / 初々
(*_~)	( ~ ~ )	\ ( ~ o ~ ) /
(*_m')	( ~ ~ )	\ ( ^ o ^ ) / 初々
( > < )	( ~ 0 ~ ) /	( ~ o ~ )
( T _ T )	/ ( ~ o ~ ) \	( ~ ? ~ )
(*_ ' '*)	( ~ ~ )	

表 2: 2 者間で逆の極性が付与された顔文字

顔文字候補とした。

### 3.2 顔文字に対する極性判定

これまでの作業により得られた顔文字候補に対して、人手により Positive、Negative、Neutral を付与する。ここでは 873 の顔文字候補に対して評価極性のアノテイトを筆者ら 2 名で行った。結果を表 1 に示す。2 者間の一致度は 0.64 (≈ (219 + 144 + 194) / 873) であった。

これに対して、2 者の間で逆の極性を付与した顔文字候補は表 2 で示すように 20 種類にも及んだ。こうした顔文字の極性を再判定するために、付与されているツイートにおいて顔文字がどのように機能しているのかを調査した。表 3 にその結果の一部を示す。多くの場合、1 つの顔文字は付与される文そのものの極性と同じであることから、一意にその極性を決定できると考えられる。しかしながら 20 種類のうちの 3 種類は、文の極性が異なるにも拘らず使用されていた。表 3 中の例のうち下 2 つの顔文字はその例である。

## 4 顔文字を適切に扱うために

実験により対象の 873 の顔文字候補に対して 219 の Positive、144 種類の Negative の極性を付与することができた。しかしながら極性が付与できたのは全体の 35% 程度であり、その多くは極性を持たない、もしくは一意に極性が付与できない顔文字候補で全体の約 45% ほど残った。またアノテイト間の付与極性の一致率が 0.64 と、決して高くない。これは言語表現に対して確立されている評価極性辞書構築法が、顔文字に対しては必ずしも適用できない可能性を示唆していると考えられる。そこで、顔文字の特徴や特有の使われ方から、より効率良く顔文字情報を扱うための枠組みについて議論する。

### 4.1 顔文字の解釈と曖昧性

顔文字が出現するメールやチャットなどの文字コミュニケーションでの対話は、送り手と受け手があるもので形

顔文字	ツイート	P	N	E
(*_~*)	119	39	0	80
( ' ' )	13	0	2	11
( T _ T )	15	0	0	15
(~_*)	78	6	9	63
(*_ ' '*)	13	1	1	11

表 3: 顔文字を含むツイート数およびツイートに対する評価極性

<sup>2</sup><http://www.twitter.com/rakutter>

式上は双方向ではあるが、直接同時にやり取りの出来る会話や電話とは異なり、基本的に一方通行の情報伝達である。そのため、顔の表情やジェスチャ、声のイントネーションや強調などの韻律が担う情報、フィラーやポーズなどのパラ言語的要素が、顔文字という他の形で表現される。

(4) もう夕寝の時間やわ (T.T) ~

(5) うちらホンマによい働くワ (^ ^)

上の例 (4) の眠気は会話ならば直接、表情やしぐさなどにより相手に伝えることができるが、文字だけでは不可能なので顔文字により図像化して伝えたものである。また例 (5) では自己賞賛の発言への照れが笑顔の顔文字によりパラ言語的要素を補足して伝達している。しかしながら通常、言語の語彙において語彙によって説明されるものと、その意味が対応関係を持つが、顔文字はそれぞれの記号列がその意味と明確な対応関係を持っているわけではない [3]。あくまでどの意味の場合にどの顔文字を使うのかは送り手の選択によってのみ決まり、受け手がどのように理解するかは分からないままである。つまり送り手が利用した顔文字を、受け手が異なった解釈でその意味を理解する可能性は否めない。こうした顔文字の解釈に関しては、小野らの研究がある [8]。彼女らは顔文字が持つ意味についての調査を行い、カテゴリ分類を行った。その中で、“(^\_^;” や “(>\_<)”、“(;;)” については回答者により「快」、「不快」の両方の解釈があったことが示され、個人または文脈により顔文字の意味が異なることが示された。先の実験によって付与された顔文字に対する極性のアノテータ間の一致率の低さは、この顔文字の解釈の個人差が原因であるといえる。また、ツイートの文脈の極性が異なるにも拘らず同じ顔文字が付与されている例もある。この例 (6a)、(6b) はどちらも同一ユーザによるものである。ツイートの一部を示す。

(6) a. (\*´´\*) あー寝れないなあ。寝れないときは無理して寝ない方が

b. キューピーがむっちゃ可愛い(\*´´\*)

#### 4.2 顔文字の分類

先の実験では、言語表現に対する評価表現辞書の構築法をそのまま顔文字に適用したが、実験結果では、およそ 45% の顔文字に極性が付与されなかった。これは顔文字のもつ意味が、個人や文脈により曖昧となることが一つの原因であると考えられるが、問題は他にも考えられる。本実験で用いた顔文字候補のうち極性が付与されなかったもので、利用したツイート上で高頻度のものに、“^^;”、“^\_^;”、“(^\_ ^;)”、“m(\_)\_m”、“( ; )” などがある。これらの顔文字は、候補全体においても同様に高頻度である。極性という枠組みで顔文字候補の分類

を考えた場合、これらを感情表現ではない (= 分類対象外もしくは Neutral クラス) と仮定することもできる。しかしながら、これらは実際の文中においては半否定、皮肉などの重要な役割を果たし、これらを捉えるとすると 2 値分類の枠組みでは限界があり、多クラス分類の検討が必要となる。

山口らは顔文字を「微笑み系エモティコン」と「冷や汗系エモティコン」という、極性ではない基準で分類している [7]。花井らはこれに従い 69 種類の顔文字の分類を試みたが、その 2 つのクラスに分類することが難しい顔文字が多くみられたため極性を導入し、さらにそれぞれの極性を次のように細かく分類した [13]。

ポジティブ感情 笑顔系、喜び系、照れ笑い系

ネガティブ感情 焦り系、落ち込み系、痛み系、泣き顔系、困り顔系

また川上らは顔文字のデータベース構築を目的として、顔文字がもつ文脈依存性を考慮し“どの感情を表しているか”、つまり唯一の分類クラスに割り当ててではなく複数の感情の軸を準備して“それぞれの感情をどの程度表しているか”という観点で分類を試みた [11]。この手法ではまず、44 個の顔文字を/笑い/泣き/怒り/焦り/驚き/その他/の 6 クラスにまず分類してそれぞれのクラスで近親性の高い顔文字を選別し 31 種類の顔文字を対象とした。次に被験者に対して顔文字が/喜び/悲しさ/怒り/楽しさ/焦り/驚き/をどの程度表しているかを 1~5 で判断させ、それぞれの顔文字の特性を評定した。結果を見ると、被験者はそれぞれのクラスの顔文字に対してほぼ正しい評定値を与えており、送り手が顔文字を利用するときは少なくともこの粒度で分類しているといえる。川上らは更に“喜び”と“楽しさ”の相関係数が極めて高く、実質的にこれらを分ける必要はない、ネガティブ感情の 4 クラスを適切に分類するのは難しいと報告した。

彼らの報告では、対象としていた顔文字が 31 個と決して多い数ではないが、手法そのものは規模に依存しておらず、識別結果のように 2 極性クラスへの分類から、焦り、驚きなどを加えた 5 クラスの分類であってもそれぞれを区別できるならば、規模を大きくしても同様の結果が得られると考えられる。また顔文字を唯一のクラスに分類するのではなく、川上らのように顔文字を複数の感情を統合的に扱う形で扱うことで、顔文字の持つ意味の緩やかさを捉えることができると考えられる。

#### 4.3 顔文字による強調、配慮

顔文字は感情のほかにも様々な役割を果たしている。井上らは発信者の感情の表現のほかに、“強調のための表現”、“きつい表現を和らげる”を挙げている [1]。また原田も同様に、顔文字と擬態語などの文字列を添えて同

時に表記すると、感情と擬態、擬声、擬音などが直接感覚的に伝えることが可能になる、と強調についても言及した。また顔文字には受け手の顔が立つよう送り手の気配りが顔文字に託されるという、配慮の機能について述べた [3]。

例えば「財布がない」と文字で表現しただけでは、送り手自身の目指す表現にはまだ情報が足りないの、より感覚的に表すために顔文字を付与する。たとえば以下のような例が考えられる。

(7) それがさ、財布がない (° °) ガーン

(8) ちょっと冒険チックやろ? (^ ^) ワクワク

この場合、顔文字と続く擬態語が持つ意味はそれぞれほぼ同じであると考えられることから、顔文字は強調の役割を果たすものと考えられる。

川上らは顔文字の感情を 5 クラスに拡充しただけでなく強調についても同様に調査し、それぞれの感情のうち典型的なものは言語表現の感情を強調することを示した [11]。井上や原田が言及した配慮には例 (3) や、次の例がある。

(9) 間違ってもうちのグループには来ないで (^ ^)

(3) の例では、「わかった」という言語表現だけでは仕方なくそう答えたのか、快く述べたのかは判断が極めて難しいが、例のような顔文字が付与されていれば、快諾していることが視覚的に理解できる。例 (9) では相手に対して「こないで」という拒否表現を用いているものの、それだけでは相手に対して強い口調になりえるので、例のような顔文字を付与することでそれが冗談であることを示している。

こうした“強調”や“配慮”は、顔文字そのものが持つ意味情報ではなく、送り手と受け手が円滑なコミュニケーションを行うために必要な、顔文字の使われ方に関する情報である。この情報の認識は、顔文字の送り手の字面の背後にある意図を捉えることであり、様々な用途で利用が可能である。

#### 4.4 顔文字を扱うタスクの設定

荒川らは、顔文字研究の今後として 4 種類の研究について述べた [4]。感情情報解析などはその中の「顔文字であるための条件研究」の中に包含される。顔文字と感情情報である、喜び/悲しさ/怒り/焦り/驚き、との対応は、辞書の構築が有効である。また未知の顔文字に対しては、田中らが取り組んでいるように [14]、統計的なアプローチの有効性が示されている。また前節で示したような、強調、配慮などの顔文字の使われ方の認識タスクもある。顔文字の感情と文の言語表現の極性との関係を推定することで、それらを識別できると考えられる。

## 5 まとめ

これまでそれほど注目されてこなかった文中の顔文字に焦点を当て、顔文字とその感情情報の辞書を作成するために言語表現の評価表現辞書と同様の手法で顔文字に極性付与を行った。顔文字の特性により単純な極性では網羅的に情報集約することが難しいことが実験から示されたため、顔文字情報を適切に扱うために顔文字の解釈、分類、使われ方について考察を行った。

今後はまず、複数の感情を統合した形の感情情報付与の有用性、効率性の検証や人手による試行、仕様策定を行う必要がある。また顔文字の極性と文の極性がどのような関係のときに、その顔文字が強調や配慮として使われるのかの事例分析を行う予定である。

## 参考文献

- [1] 井上みづほ, 藤巻美菜子, 石崎俊. 電子メール文における感情表現の解析システムについて - 感情表現の収集・分類・解析 -. 電子情報通信学会技術研究報告 TL96-11, pp. 1-8, 1997.
- [2] 井上逸兵. ネット社会の若者言葉. 月刊言語, 第 35 巻, pp. 60-67. 大修館書店, 2006.
- [3] 原田登美. 「顔文字」による日本語の円滑なコミュニケーション: 「配慮」と「ポライトネス」の表現機能. 言語と文化 (甲南大学), Vol. 8, pp. 205-224, 2004.
- [4] 荒川歩. 顔文字研究の原状と展望: 非言語コミュニケーション研究の視点から. 心理学評論, Vol. 50, No. 4, pp. 261-370, 2008.
- [5] 荒川歩, 鈴木直人. 謝罪文に付与された顔文字が受け手の感情に与える効果. 対人社会心理学研究, Vol. 4, pp. 128-133, 2004.
- [6] 高村大也, 乾孝司, 奥村学. 隠れ変数モデルによる複数語表現の感情極性分類. 情報処理学会論文誌ジャーナル, Vol. 47, No. 11, pp. 3021-3031, 2006.
- [7] 山口英彦, 城仁士. 電子コミュニティにおけるエモティコンの役割. 神戸大学発達科学部研究紀要, Vol. 8, No. 1, pp. 131-145, 2000.
- [8] 小野聡子, 原田知沙, 徳田克己. 大学生におけるメールの利用について i - 顔文字の意味の認識 -. 日本教育心理学会大 45 回総合発表論文集, p. 267, 2003.
- [9] 小林のぞみ, 乾健太郎, 松本裕治. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 2, pp. 203-222, 2005.
- [10] 角野清久, 永井由佳里. 顔文字を使用した文章によるコミュニケーションでの受け手の解釈の傾向. 日本認知科学会第 24 回大会, pp. 118-119, 2007.
- [11] 川上正浩. 顔文字が表す感情と強調に関するデータベース. Technical report, 大阪樟蔭女子大学人間科学研究紀要, 2008.
- [12] 竹原卓真. 世代の違いによる顔文字の感情伝達効果. 日本顔学会誌, Vol. 7, pp. 37-46, 2007.
- [13] 花井友美, 小口孝司. E メールの交換過程における感情用言の出現パターン. 社会心理学研究, Vol. 24, No. 2, pp. 131-139, 2008.
- [14] 田中裕紀, 高村大也, 奥村学. 文字ベースのコミュニケーションにおける顔文字に関する研究. 第 7 回言語処理学会年次大会, 2004.