顔文字の原形抽出

奥村 紀之
香川高等専門学校 情報工学科
okumura@di.kagawa-nct.ac.jp

1 はじめに

インターネットの発展により，文字を基本としたコミュニケーションが定着してきている．一方で，文字のみの情報のやり取りでは書き手の意図が適切に読み手に伝わらず齟齬が生じ，思わぬトラブルに発展することがある．本研究で対象としている顔文字は，文字のみのやり取りの中で，書き手の表情や感情，周囲の状況などを伝える補足情報として付加される．

日本語はハイコンテクストな言語であると指摘されている [3]．ハイコンテクストな言語とは，言語として明確に表現された内容よりも，明示的に言葉に示していないにも関わらず，相手に理解されるであろうと期待している情報量の方が豊かな言語である．これに対し，ローコンテクストな言語とは，言語として明確に表現した以上のことは基本的に伝達されない言語である．

Hall の指摘を否定する向きもあるが [2]，我々の日常生活の中で，対面で行われるコミュニケーションでは，親しい間柄であればあるほど多くを語らず，その場の雰囲気で察することを要求されることが多いことに気づく．特に日本人同士では，沈黙は金，雄弁は銀と言われるように，多くを語らず，状況に応じて適切に相手の心境を想定し，行動することを要求される．

このような文化的背景と文字を基本としたオンラインコミュニケーションの相性は非常に悪く，不用意な発言からブログにおける炎上などのトラブルに発展することがある．これは，投稿者の表情など言葉以外の情報が適切に読み手に伝わらず，文字通りの解釈をしてしまったり，言葉の奥にある情報を読み違えたりしたため起きる問題である．

そこで，文字を基本としたオンラインコミュニケーションをより我々の価値観に合わせたものとするため，自然発生的に顔文字の使用が盛んになっている．顔文字は，スマイリーと呼ばれる :-) のような正面から見て 90 度回転したようなものに加え，(^_^) のように正面を向いたものが存在している．スマイリーは主にアルファベットを使用する言語圏で使用され，その数はおよそ 100 種〜500 種程度と少数である．

一方，正面を向いた顔文字は日本語圏で特に使用頻度が高く，その数は現在確認できているだけでも 10 万種以上とスマイリーと比較しても圧倒的に多い．そのため，Twitter やブログなど顔文字が使用される頻度が高い文書を適切に解釈するためには，日本語圏で使用されている顔文字に関する大規模な辞書が必要となる．

本研究では，Web から収集した 69,026 種のうち 22,000 種の顔文字に対し，顔文字の原形に加え，付属しているパーツ，コメント，顔文字から想定される感情といった情報を付与し，大規模な顔文字のタグ付き辞書を構築している．本稿では，言語処理の観点から利用可能な，顔文字の原形に関する調査報告を行う．

2 関連研究

顔文字の原形に直接的に関係する研究はないが，Ptaszynski らが開発している CAO システム [5] では，目-口-目の並び (Triplet) に着目して顔文字の抽出を行っている．Ptaszynski らのシステムでは，データベースの規模としておよそ 1 万種の顔文字に対応しており，さらに自動拡張によっておよそ 300 万種に対応可能であるため，顔文字の研究の中では非常に規模の大きいものである [6]．同様に，機械学習によって顔文字を抽出する研究には Tanaka らの研究もある [8]．

文中に現れる顔文字の抽出方法としては，Bedrick らの手法がある．顔文字の抽出手法として正規表現が一般的に用いられるが，Bedrick らは，HMM によって記号列を抽出し，PCFG に基づく評価法によって顔文字の候補を選別する手法を提案している [1]．Bedrick らの研究において使用されている PCFG のルールから，顔文字は”対称性”を持つことが重要であると考えられ，本稿における顔文字の原形抽出においても密接に関係している．

本稿で目的としているような言語処理のための顔文字に関連する辞書には，Sato が構築している形態素解析器 MeCab で利用可能な mecab-ipadic-NEologd がある [7]．MeCab の設計方針として，辞書と解析器の分離があり，活用形等の情報は全て辞書に記載されている．Sato の開発している mecab-ipadic-NEologd では，顔文字も形態素解析用の情報として記載されているが，原形の項目については，対象としている顔文字そのもの，あるいは軽微な補正を加えたものに留められている．

3 顔文字

顔文字は，いくつかの記号または文字を組み合わせることによって表現される文字列 (記号列) であり，言語としての意味を表現するというよりも，ジェスチャーなどのボディランゲージや顔 (表情) を認識できるという特徴を持つ．さらに，顔としての表情を表現するという基本的な機能に加えて，台詞やオノマトペが複合されることで，感情や状況の表現力が高められた亜種が増加している (Ex.(^−^)_θ お薬ですお大事に，(i_i)＼(^−^) ヨシヨシ)．

顔文字が爆発的に普及し，現在も新たな顔文字が日々生み出されている状況において，顔文字を適切に解析し，顔文字の持つ情報を解釈するためには，大規模な顔文字の辞書が必要となる．顔文字の辞書に持たせる情報としては，顔文字とそれに対する読み (あるいはラベル)，含有する感情など，多岐にわたる．本稿では特に，顔文字の原形に着目し，顔文字のパーツ単位での解析を目的とした分析を行う．

3.1 顔文字の単位

顔文字の中には，(^_^) のような単純なものから，「蚊がいるぞ丶 (´⊇`) ノはぁ〜〜(*人*) パン！ 」のように複数の顔文字が組み合わされ，台詞やオノマトペが複合されることにより，一種のストーリー性を持つものまで多種多様なものが存在している．一方で，顔文字とは何かといった議論が深くなされたことはなく，研究者の間でも顔文字の単位が統一されていない問題がある．

本稿で対象としている顔文字は，以下の 5 つのうち 1 つ以上の項目に合致する記号列を顔文字として扱っている．

1. 顔を表現すると思われる記号列
2. 身体の一部 (腕や足など) を表すと思われる記号列
3. 顔文字が発していると考えられる台詞の列
4. 状況を表すためのオノマトペ
5. 複数の顔文字が出現するような記号列

これらの条件を満たす最長の記号列を顔文字の単位とし，辞書の構築と原形抽出を行う．

3.2 顔文字の原形

Ptaszynski らのシステムでは，目-口-目の Triplet を最小単位として解析し，感情抽出や顔文字の検出を行っている．しかし，顔文字の中には口を持たず，目-鼻-目のようなタイプや目-目のタイプなど，基本単位とする Triplet に合致しないものも多数存在している．Ptaszynski らは／(ノ ω ；) ＼シクシクのような顔文字では，ノ ω ；を Triplet として抽出している．この場合，；について涙を流している目だと解釈し，ノについては，目を覆っている腕 (手) と解釈することが自然だろう．

また，Bedrick らの手法にも見られるように，顔文字の対称性という観点から評価すると，Triplet の定義は厳密ではないと考えられる．Sato の mecab-ipadic-NEologd においても，顔文字の原形が登録されているが，関連研究で述べたとおり，軽微な補正に留められており，対称性は考慮されていない．

本稿では顔文字の原形を，Ptaszynski らの Triplet に加え，輪郭を表現する文字列を加えたものとして定義する．顔文字の原形抽出は，表 1 のルールに従う．

顔文字による表現は多様で，表 1 の方針のみでは原形を定義できないものも存在している．例えば，( ) という顔文字は後頭部を表現している．このような顔文字については，目や口を補完することが困難であるため，後ろ姿のまま抽出するなど例外的な扱いとしている．対応する半角文字が存在しない例は，輪郭，目，口 (鼻) 以外のパーツは除去するというルールを適用後，ω を変換しない例として説明している．

3.3 顔文字の原形を定義する意味

顔文字を言語処理的な観点から処理したい場合，単純な記号列として扱うよりも，原形と活用形として処理した方が都合が良い場合がある．特に，日々増え続ける顔文字に対して逐次原形抽出を行うことは非常に手間がかかるため，原形を推定するモデルが存在すれ

表 1: 原形抽出のルール

顔文字	原形	ルール
(^␣^)	(^^)	スペースは全て削除する
\(^_^)/	(^_^)	輪郭，目，口 (鼻) 以外のパーツは除去する
(^ __ ^)	(^_^)	元の顔文字が全角で表現されている場合は，半角に変換する
(’・ω・‘)	(・ω・)	該当する半角文字が存在しない場合は全角のままとする (この場合はωが対象)
(T_T)/~~~	(T_T)	大文字小文字の変換は行わない (例えば (t_t) のようにはしない)
(>_-)	(>_<), (-_-)	左右非対称の場合は，対称となるよう複数の原形を抽出する
(ノω;)	(;ω;)	腕などで目が隠されている場合は，他方の目に合わせて補完する
(^_^;	(^_^)	輪郭が一方にしかない場合は，対応する輪郭を補完する
(￣∀(・。・;)	(￣∀￣), (・。・)	複数の顔文字がある場合は，それぞれの原形を抽出する
)^^)	(^^)	輪郭が非対称である場合は，膨らみを外側に向けるよう補正する

ば，原形以外のパーツを活用形と見なして分類することが可能となる．

原形を推定するモデルに関しては本稿では触れないが，Ptaszynski らの CAO システムのように Triplet を抽出するステップを利用する方法など，既存の手法でも対応が可能である．また，顔文字の活用形に関しては，パーツごとの意味を解釈するための実験を別途行っており [4]，より詳細に顔文字を解釈するシステムが期待される．

4 アノテーション

アノテーションは，表 1 に示した基準に則り，6 名の被験者により実施している．アノテーション作業開始時点 (2015 年 10 月) で 69,026 種の顔文字を収集していたが，本稿執筆時点で完了しているアノテーションは 22,000 種，複数の顔を有する顔文字に対するアノテーションを考慮すると 37,799 種であり，現在も引き続き作業を進めている．そのため，本稿では，執筆時点までに完了しているアノテーションに基づき検証している．

なお，今後の分析のため顔文字の原形抽出の他に，付属しているパーツの特性を検証し，顔文字の活用形に関する情報として表 2 のものを付与しているが，本稿では言及しない．

5 結果

表 3 にアノテーションにより抽出できた顔文字の原形の例を示す．頻度は 37,799 種のアノテーション済みの顔文字の中で，当該の原形を持つ顔文字の個数を示している．

抽出できた顔文字の原形は 3,071 種あり，そのうち 1,183 種は頻度が 1 であった．頻度が 1 であったものは，他の顔文字と共通の原形を有しない特徴的なものであるため，その例を表 4 に示す．

表 2: 顔文字へのタグ付け

台詞区切り	左台詞	左腕	左パーツ
左耳	左輪郭	左頬	左眉
左目	額	鼻	口
右目	右眉	右頬	右輪郭
右耳	右腕	両腕	右パーツ
右台詞	オノマトペ	繰り返し	原形
半角変換	顔文字判断		

表 3: 抽出した顔文字の原形の例

顔文字の原形	頻度	顔文字の原形	頻度
(･_･)	653	(‥)	353
(___)	443	(0￥0)	351
(･ω･)	439	(･エ･)	343
(-_-)	438	(^^)	338
(>_<)	432	(￣ｗ￣)	332
(Φｗ Φ)	398	(--)	253
(○|○)	370	(゜○゜)	245
(^-^)	362	(゜д゜)	232

表 4: 頻度が 1 であった顔文字と原形

顔文字	原形
ヽ(ΘπΘ)ノジュルジュル	(ΘπΘ)
[ヽΘ´]]](中略) ＝３ マテェー！！	[ヽΘ´]
(━━━┻━━━○━━━┻━━━)	(┬○┬)
ヽ(+Θ+)ノ・・・キュゥ	(+Θ+)
(前略)(ドカーン))))))☆(/.x)/アレー	(x.x)

6 考察

表 3 に示したとおり，顔文字の原形を抽出すると数百もの顔文字が，共通する原形を有していることが分かる．顔文字は言語とは異なり，何らかの法則に基づいて変化をするわけではなく，必要に応じてありとあらゆる記号が顔文字を拡張することになる．そのため，共通する原形を有する顔文字に付与されているパーツに着目して，表情や感情など顔文字によって表現されている情報がどのように変化するかに着目した分析が必要となる．

全角を含む顔文字を可能な限り半角で表現し多様性を抑制することによって，多くの顔文字が共通した原形を有することが分かる．一方で，原形の頻度が 1 の顔文字に着目すると，アノテーションを施した 37,799 種のうち 1,183 種，比率にして約 3%とその割合は少ない．今後，新しい顔文字を抽出したり，その原形を自動抽出することを考えると，95%以上の顔文字が，他の顔文字と原形を共有しているはずだという前提での解析をしても大きく問題にはならない．

しかし，本稿で検討できている顔文字は Shift-JIS の文字コードで表現可能なものがほぼすべてであり，例えば，図 1 のような UTF-8 で使用可能な特殊文字を組み合わせた顔文字については対応ができていない [1]．

ლ(╹◡╹ლ)	(｡◕‿‿◕｡)	(￣◡￣)	(*→‿←*)	(๑╹◡╹๑)
(๑→‿←)ｷｭｯ♡	(｡◕ω◕｡)	(･Θ･)	(ù_, úo)ﾌｯ	(´｡◕ω◕｡)
(｡ì _ í｡)	(｡◕ ∀ ◕｡)	わこちゅー(U◕ﻬ◕)ﾉ	(๑ˆ∀ˆ๑)	(｡╹ლ╹)
(｡í _ ì｡)	(❀◕‿◕)こんばんわぁ			

図 1: UTF-8 の特殊文字を利用した顔文字

7 おわりに

本稿では，大規模な顔文字に関する辞書について，特に顔文字の原形について述べた．インターネット上に大量に流通している顔文字は，一見するとどれもが特徴的な顔文字と考えられるが，原形を定義することによって，顔文字をいくつかのグループに分類できることが分かった．

今後は，同時並行で進めている顔文字の各パーツに関して原形との関係を抽出し，感情の遷移，パーツごとの特徴などを詳細に分析していきたい．

謝辞

本研究は JSPS 科研費 15K21592 の助成を受けたものです。

[1] Unicode 顔文字：http://july.mydns.jp/

参考文献

[1] Steven Bedrick, Russell Beckley, Brian Roark, and Richard Sproat. Robust kaomoji detection in twitter. In *Proceedings of the Second Workshop on Language in Social Media*, p. 56–64, Montréal, Canada, Jun. 2012. Association for Computational Linguistics.

[2] Peter W. Cardon. A critique of hall's contexting model: A meta-analysis of literature on intercultural business and technical communication. *Journal of Business and Technical Communication*, Vol. 22, p. 399–428, Oct. 2008.

[3] Edward T. Hall. *Beyond Culture*. Anchor Books, 1976.

[4] Chika Onishi and Noriyuki Okumura. An investigation of the usage of kaomoji for emotions judgment and kaomoji recommendation. In *The 13th IASTED International Conference on Artificial Intelligence and Applications AIA2014*, p. 334–341, Feb. 2014.

[5] Michal Ptaszynski, Jacek Maciejewski, Pawel Dybala, Rafal Rzepka, and Kenji Araki. Cao: A fully automatic emoticon analysis system based on theory of kinesics. *Affective Computing, IEEE Transactions on*, Vol. 1, No. 1, p. 46–59, Jan. 2010.

[6] Michal Ptaszynski, Jacek Maciejewski, Pawel Dybala, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi. *Speech, Image, and Language Processing for Human Computer Interaction - Science of Emoticons: Research Framework and State of the Art in Analysis of kaomoji-type Emoticons*. IGI Global, Jan. 2012.

[7] Toshinori Sato. Neologism dictionary based on the language resources on the web for mecab, 2015.

[8] Yuki Tanaka, Hiroya Takamura, and Manabu Okumura. Extraction and classification of facemarks. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, IUI '05, pp. 28–34, New York, NY, USA, 2005. ACM.