3H3-OS-17a-1

言語解析のための大規模顔文字辞書

A Large Scale Dictionary of Kaomoji for Natural Language Processing

奥村紀之
Noriyuki Okumura

明石工業高等専門学校 電気情報工学科
National Institute of Technology, Akashi College, Department of Electrical and Computer Engineering

In this paper, we describe the detail of a large scale dictionary of Kaomoji for natural language processing. Kaomoji are widely used in customer generated media, howerver, they have a large number of expression that express writer's emotions, conditions, situatlons ,and so on. We build a Kaomoji dictionary thar includes their original form and their parts such as eyes, nose, mouth, border, ears, etc. In recent years, Kaomoji have not only their "face" but their "comment" and/or "onomatopoeia". We also extract these linguistic information and give them to our Kaomoji dictionary.

1. はじめに

本稿では，オンラインコミュニケーションを円滑に進めるために広く活用されている顔文字に関する大規模な辞書構築に関して報告する．顔文字は主として 2 種類のパターンに分類できる．一つ目は，:-) のように正面から見て 90 度回転しているアメリカ式の顔文字，二つ目は，(^o^) のように正面を向いた日本式の顔文字である．アメリカ式の顔文字は，総数で 500 種程度と少数で有り，変化の多様性が見られない．一方で，日本式の顔文字はすでに 10 万種以上確認されており，今後も益々増加していく物と考えられる．本稿では特に，正面を向いた日本式の顔文字に関しておよそ 4 万種規模の大規模な辞書を構築している．

2. 関連研究

顔文字の辞書の利用方法として，オンライン資源を利用することや，携帯電話などに標準で搭載されている辞書を参照することが考えられる．オンライン資源では顔文字屋 *1 などがあり，大量の顔文字とその読みが付与された辞書を参照することができる．携帯電話などに搭載されている顔文字の辞書は，"かおもじ"と入力し仮名漢字変換を利用することで簡単に参照できる．しかしながら，これらの辞書は顔文字と読みの関係しか定義されておらず，顔文字の詳細な分析はできない．

顔文字の辞書構築に関する先行研究としては，筆者らが進めている [奥村 14] がある．顔文字を構成するいくつかのパーツに着目した分析を進めてきているが，顔文字の原形に着目した分析ではなく，統合的，統計的な扱いが困難であった．そこで，[奥村 16] にて顔文字の原形を定義し，この問題を解消している．

一方，佐々木らは，文章中に含まれている顔文字 (アスキーアート) を正規化処理の一部分として削除している [佐々木 13]．佐々木らが指摘しているように，顔文字の表記法に関して正書法では考慮されていないが，我々が提案している顔文字の原形抽出など，顔文字の標準化機構を考慮することで，顔文字の多様性を抑制し，統計的な扱いも可能となる．特に，文章から抽出したい情報が文構造など言語学的な要素ではなく，感情や個性，価値観といった情報である場合，正規化処理として顔文字などの文字列を一律で削除してしまうことには問題があると考えられる．そのため，本稿で提案するような大規模な顔文字の辞書を参照することで，顔文字の情報も含めた正規化処理等も考慮すべきであろう．

連絡先: 奥村紀之，明石工業高等専門学校電気情報工学科，兵庫県明石市魚住町西岡 679-3，078-946-6017，okumura@akashi.ac.jp

*1 http://kaomojiya.com

3. アノテーション

本稿では大規模な顔文字の辞書を構築するため，様々な観点から顔文字を分解し，タグを付与している．以下，各タグの分類と詳細について解説する．

3.1 原形

顔文字を形態素解析の観点から利用しやすい形式に変換したり，顔文字に付与されている様々なパーツに意味づけを行うためには，顔文字の原形を適切に定義する必要がある．本項目に関しては既に [奥村 16] にて報告済みであるが，本稿では顔文字の原形として，顔文字に含まれている目，口 (鼻)，輪郭を基本構成要素として抽出している．表 1 に原形の例を示す．

表 1: 顔文字の原形

顔文字	原形	顔文字	原形
(´-ω-`)	(-ω-)	”(@´▽` @)ノ””	(´▽`)

3.2 目のパーツ

顔文字としての最低限の構成要素として，本稿では目を重視している．すなわち，後ろ姿など，顔かどうか判別しづらいものや，orz などの顔とは一概に言いがたいような顔文字に関しては別項目として取り扱うこととしている．顔文字には，輪郭や鼻が存在しないもの，例えば，^^などの顔文字が存在している．そのため，アノテータが目を認識できないような顔文字に関しては除外して辞書を構築している．表 2 に顔文字から抽出した目の例を示す．

3.3 口・鼻

顔文字の構成要素として目の次に多く検出されたものが口である．その他，口が存在しない代わりに鼻が付与されている

表 2: 顔文字の原形を構成する目

顔文字	左右の目	顔文字	左右の目
(－ｏ－)	－ －	(゜～゜)	゜ ゜

顔文字も存在し，アノテータ間で双方とも同程度に重要だという共通見解が得られたため，口，あるいは鼻を顔文字を構成する基本要素として原形にも含め，抽出している．表 3 に口や鼻の例を示す．

表 3: 顔文字の原形を構成する口・鼻

顔文字	口	顔文字	鼻
(((゜Д゜)))	Д	(*⁻ ^^⁻ *)	^^
(’ー’)	－	(・@@・)	@@

3.4 輪郭

顔文字の原形を構成する基本要素として，輪郭を抽出している．輪郭は通常，丸括弧 ( ) によって表現されることが多く，他の構成要素と比較すると分かりやすい構成パーツである．しかし，丸括弧が使用されず，例えば女性の髪型を表すような輪郭が含まれる場合，他の文字列との比較によって顔文字が分断されてしまう恐れがある．そこで，表 4 に示すような輪郭を抽出している．

表 4: 顔文字の原形を構成する輪郭

顔文字	輪郭	顔文字	輪郭
(・∀・)	( )	州・_・州	州 州

3.5 頬・耳・額

顔文字を構成する基本要素以外のパーツとして，まず顔そのものに関連する要素が頬・耳・額である．これらの要素は，必ずしも顔文字に存在しているということはなく原形には含めていない．つまり，原形以外パーツは，原形に対して何らかの変化を与える (感情の変化，状態の変化など) ための補足パーツとして位置づけている．表 5 に頬・耳・額のそれぞれの例を示す．

3.6 腕

顔文字の表現を豊かにするため，腕を表すパーツが付与されることもある．腕の項目は/などで表現されることが多く，主として動作を表すものが多い．腕のパーツは，左右両方が同時に出現するものもあれば，片腕だけしかないもの，両腕を一つの文字で表現するものなど多様な表現形式を持つ．表 6 に腕のパーツの例を示す．

3.7 その他のパーツ

顔文字を構成する要素として，顔や身体の一部を表すパーツ以外にも様々なパーツが付与されることがある．特に，状況を詳しく表現したい場合や，感情を強調したい場合に多く使用されている．これらのパーツは，顔文字に含めるべきではない，すなわち，顔文字を構成する要素は顔に限定するという議論もあるが，本稿では今後も増加すると予測される顔文字の表現に対して，様々な角度から自動的な判定を行うための基本的な分析を進めることを目的としており，周辺に存在する顔や身体以外のパーツについても抽出している．なお，これらのパーツは顔文字に対して右側に置かれるか左側に置かれるかによっても扱いやすさに影響が出るため，左右の別も記載している．特に，文章中に出現する顔文字とそれ以外の文章の部分に分割するためには，特に顔の左側に出現するパーツの傾向を知らなければ，文章と顔文字の境界が曖昧になり，適切な処理が難しくなる．表 7 に例を示す．表の例では，旦がお茶を表現している．

表 5: 頬・耳・額のパーツの例

顔文字	頬	顔文字	耳	顔文字	額
(*^^*)	* *	⊂(・(ェ)・)⊃	⊂ ⊃	(－゛－)	゛
(゜〇゜;)	;	∈(・●・)∋	∈ ∋	－	－

表 6: 腕のパーツ

顔文字	左右の腕	顔文字	両腕
┗|⌒ｏ⌒|┛	┗ ┛	(－人－)	人

表 7: 顔や身体以外のパーツ

顔文字	パーツの例
(－_－) 旦~	旦~

3.8 台詞・オノマトペ

顔文字単体では表情や動作しか表現することができないが，顔文字に台詞やオノマトペが付与されることで，表現力が向上する．オノマトペに関しては，オノマトペ辞典などを参照することで比較的容易に検出可能ではあるが，台詞に関してはありとあらゆる言語表現が台詞として付与することが可能であるため，全ての台詞を抽出することはできない．一方で，台詞の付与された顔文字が存在することも事実であり，この情報は簡単には無視できない．そこで，台詞の一例も顔文字の分析対象として抽出している．表 8 に簡単な例を示す．

3.9 その他の構成要素

これまでに述べた顔文字のパーツ以外にも，特徴的な顔文字の要素が存在している．例えば，)))))) など，括弧を連続させることによって震えている様子を表現したり，何かを投げつけているような残像を表現したりする．しかし，どの顔文字でも同様に同じ数だけ連続させているわけではなく，一般に連続するパーツの数は顔文字の表現に依存している．そのため，連続している項目が含まれていれば，その項目を抽出し，タグを付与している．

また，顔文字の，特に原形の多様性を抑制するため，全角文字と半角文字の変換操作を行っている．これは，例えば，^という記号は目として使用されることが多いが，＾のように全角文字で表現される場合がある．この場合，(^^) という顔文字と，（＾＾）という顔文字が大きさの違い以外に差がないにもかかわらず別々の原形として抽出されてしまう．この問題を避けるため，顔文字を構成する要素に含まれている全角文字の中で，それに対応する半角文字が存在していれば，半角での表現に変換している．この作業を要する顔文字か否かを示す情報を付与している．

表 8: 台詞・オノマトペ

顔文字	台詞	顔文字	オノマトペ
(’▽’)ノ やぁ	やぁ	”ρ(’-’*) ピッ	ピッ

顔文字にアノテーションする際，単一の顔しか出現しない顔文字であれば作業は容易であるが，複数の顔が存在するような顔文字も非常に多く観測されている．そのため，アノテーション作業では複数の顔が含まれている顔文字の場合，左側から出現順に原形抽出を行い，着目している顔文字に対するパーツとして，他の顔文字を位置づけている．

4. まとめと展望

本稿執筆時点でアノテーションが完了している顔文字の種類は 38,641 種あり，Web から取得したおよそ 7 万種の顔文字の半数に対して処理が完了している．今後，本稿で述べた基準に則した顔文字のアノテーション作業を完成させるとともに，顔文字の原形を自動抽出する手法や感情解析など，顔文字を利用する様々なアプリケーションの開発を進めていく．

特に，顔文字に対する時制の付与に関しては，[Onishi 14] で報告しているように大規模なテキストデータの解析が必要となる．現在構築している顔文字の辞書との比較検討を進めながら，顔文字の持つ時制情報についても自動的に付与していきたい．

参考文献

[奥村 14] 奥村紀之，大西智佳：顔文字に含まれる感情成分の分析と感情極性辞書の構築，言語処理学会第 20 回年次大会，P7-18，(2014)

[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，P1-1，(2016)

[佐々木 13] 佐々木彬，水野淳太，岡崎直観，乾健太郎：機械学習に基づくマイクロブログ上のテキストの正規化，第 27 回人工知能学会全国大会，4B1-4，(2013)

[Onishi 14] Chika Onishi and Noriyuki Okumura：An Investigation of the Usage of KAOMOJI for Emotions Judgment and KAOMOJI Recommendation，The 13th IASTED International Conference on Artificial Intelligence and Applications AIA2014，816-014，pp.334-341，(2014)