Twitter を対象とした絵文字の用例の自動分類

友山 孝広　　延澤 志保
東京都市大学知識工学部情報科学科

1. はじめに

近年普及しているソーシャル・ネットワーキング・サービスのようなテキストによるコミュニケーションでは，感情や状況を表現するために顔文字や絵文字が用いられている．本稿では，絵文字と共起する語を用いて絵文字の意味や使われ方を特定する手法を提案する．

2. 絵文字

絵文字とは記号を用いて人の表情などのさまざまな事物を表現したものである．絵文字は，その用途によって，装飾的，意味的，機能的の 3 種類の用法に分類できる[1]．絵文字は絵であるため，使用する人，箇所によって意味が変わるという曖昧性があるため，日々絵文字の意味が変化していると考えられる[2]．

3. 提案手法

本稿では，絵文字の意味属性の自動抽出のため，Twitter から絵文字の使用例を収集し，絵文字を意味クラスタに自動的に分類する手法を提案する．

3.1 ツイート群の収集

絵文字使用例とするツイート群は絵文字ごとに収集する．Twitter は絵文字の検索に対応していないが，異体字セレクタと呼ばれる文字コードをキーワードに検索をすることで，異体字セレクタに対応している絵文字を含むツイートが取得可能である．

3.2 絵文字の意味クラスタ候補の抽出

収集したツイート群の文から，キーワードとして名詞，動詞，形容詞，感動詞を抽出する(表 1)．抽出されたキーワードのうち，別の感動詞のキーワードと半分以上前方一致する場合は略語であると考え，これらを同じキーワードとして統合する．例えば「おは」は「おはよう」に統合される．抽出されたキーワードのうち，出現頻度の高い上位 10 個のキーワードを意味クラスタ候補とする．

表 1 絵文字を含む文と抽出キーワードの例

絵文字を含む文	抽出されるキーワード
おはようございます☀	おはよう
関東は晴れてます☀	関東，晴れ
そんなことを思う朝☀	こと，思う，朝
今日もいい天気☀	今日，いい，天気

3.3 意味クラスタの推定

ツイート群に含まれるすべてのツイートを，ツイートが含むキーワードに応じて意味クラスタ候補に振分ける．各ツイートはひとつの意味クラスタにのみ属するものとし，キーワードを複数含む文は出現頻度が高い意味クラスタ候補に振分ける(図 1)．

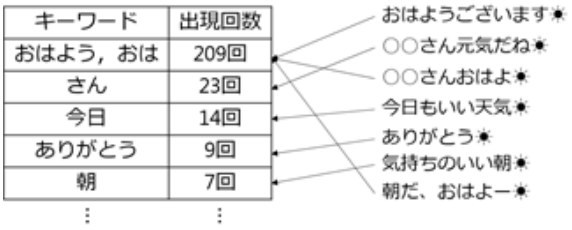

図 1 キーワードへのツイート振分け例

10 個の意味クラスタ候補のうち，振分けられたツイートの数が少ないものは，他の意味クラスタと意味的に同一である可能性が高い．そのため本稿では，出現回数に対する振分けたツイートの割合が 25%以上のものを絵文字に対する意味クラスタとする(図 2)．

キーワード	出現回数	該当する文
おはよう，おは	209回	209文
さん	23回	3文
今日	14回	5文
ありがとう	9回	6文
朝	7回	3文

図 2 意味クラスタ選択の例

4. 実験結果

2016 年 12 月 13 日に投稿されたツイートから「☀」が含まれる 272 文について提案手法を適用して実験し，本手法による分類と，予め手作業で分類した正解の分類がどれだけ一致しているかで評価をした(図 3)．

	分類名	該当する文	全ツイート中の割合
手法による分類	おはよう，おは	209文	77%
	ありがとう	6文	2.2%
	今日	5文	1.8%
	朝	3文	1.1%
正解の分類	おはよう	234文	86%
	ありがとう	10文	3.7%
	晴れ	4文	1.5%
	朝	3文	1.1%

図 3 評価実験の結果

5. まとめ

Twitter でも絵文字が複数の意味で使用されていることが確認でき，提案手法の有効性が確認できた．

参考文献

[1] 萩原正人，水野貴明，“モバイル検索システムのための絵文字に対する意味解析，” 言語処理学会第 16 回年次大会，pp.567-570, 2010.

[2] 山本千尋，別所克人，内山俊郎，内山匡，“絵文字を考慮したテキスト解析の研究，” 情報処理学会第 72 回全国大会，vol.2, pp.49-50, 2010.