7P-02

用法に着目した文中の絵文字の意味推定

高本 健太* 町田 翔* 延澤 志保*
*東京都市大学知識工学部 *東京都市大学大学院工学研究科

1 はじめに

Twitter や LINE，Facebook などのソーシャル・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ニュケーションでは，テキスト以外にも感情や状況を表すために絵文字や顔文字が用いられている．絵文字は文中で意味を持つので，文を構成する要素としてテキスト解析結果に反映する必要がある．絵文字はそのままでは通常の単語と同様に扱うことができないため，絵文字を通常の単語に置き換える必要があり，置き換えるためには，使われている絵文字の意味が分からなければならない．そこで本稿では，絵文字の用法に着目し，絵文字と共起する語を元に絵文字の意味を推定する手法を提案する．

2 絵文字の分類と意味推定

絵文字は，さまざまな用途で幅広く利用されており，その用途によって表 1 のように装飾的，意味的，機能的の 3 種類の用法に分類できる [1]．このうち装飾的絵文字と意味的絵文字の 2 種類は，文中でなんらかの意味内容を有している．絵文字はそのままでは通常の単語と同様に扱うことができないため，絵文字は文処理に当たって無視されることが一般的である．しかし，これらの絵文字は内容の強調であることも多く，また意味的絵文字はこれが無視されることで文中の重要な意味内容が抜け落ちることにもなり，文の理解において絵文字を無視することは好ましくない．

表 1: 用法による絵文字の分類

タイプ	使用法	例
装飾的	装飾目的で使用	打ち上げでビール🍺
意味的	単語の代替	この🍺は美味しい
機能的	近隣の要素に情報を与える	🅾トップ

絵文字の意味推定を行うため，友山らは絵文字の意味内容の抽出および分類を行う手法を提案している [2]．友山らの手法ではまず，絵文字を含むツイートの名詞，動詞，形容詞，感動詞をキーワードとして抽出する．次に，抽出したキーワードの出現頻度の高い上位 10 個を絵文字の用例名候補とし，絵文字の用例名候補を含む文をそれぞれに分類する (図 1)．絵文字の用例名候補を複数含む文の場合は，出現頻度の高い単語の用例名候補に分類する．友山らは，絵文字の用例名候補の中で分類された文が多い 3 つのグループを最終的な用例として出力している．しかし，絵文字の用例名候補を複数含んでいるツイートの場合，絵文字がツイートの中で出現頻度の低い単語の用例として使われていたとしても，出現頻度の高い単語の用例に分類してしまう問題がある．

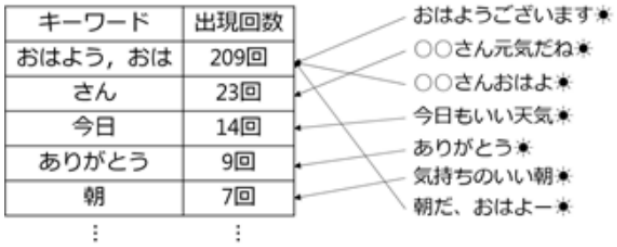

図 1: キーワードへのツイート分類の例 [2]

3 提案手法

本稿では，文の要素となっている絵文字をテキスト解析結果に反映できるようにするために，絵文字と共起する語を利用して絵文字の意味を推定する手法を提案する．

3.1 教師データの作成

教師データを作成するため，絵文字を含むツイートを収集する．単語の後ろや文末に付くことで単語や文を装飾している絵文字は，装飾している対象の意味と同じ意味を持つので，絵文字によって装飾されている単語を手がかりにして絵文字の意味を推定する．例えば「早くビール🍺飲みたい！」というツイートでは，ビールという単語を絵文字「🍺」が装飾しているので，このツイートでの「🍺」の意味は「ビール」と推定できる (表 2)．このようにして，収集したツイート中の絵文字すべてに対して意味付けを行う．

ここで，同じ意味の絵文字を含むツイート群それぞれに出現する名詞，動詞，形容詞，感動詞をその絵文字の周辺語と定義する．周辺語の TF-IDF 値を求めたところ，表 3 のような結果が得られた．TF-IDF 法とは単語の現れ方によって重要度を評価する手法であり，意味ごとの単語の傾向を調べることが出来る．表 3 を見ると，似た単語が出ても意味ごとに共起する単語の傾向がわかれている．意味ごとのツイート数と周辺語の出現総数を表 4 に

Usage-Based Disambiguation of Emoticons Embedded in Sentences
Kenta Takamoto*, Sho Machida*, and Shiho H.Nobesawa*.
* Faculty of Knowledge Engineering, Tokyo City University
* Graduate School of Engineering, Tokyo City University

表 2: 意味の推定の例

意味内容	ツイート文
ビール	早くビール🍺飲みたい！ 座るなりビールを注文🍺
飲む	今度一緒に飲もう🍺 やはり飲みは楽しいもんだ🍺
忘年会	忘年会🍺腹いっぱい食いましたね！ 今日は忘年会でした🍺
飲み会	バイトの飲み会楽しかった🍺 今日の飲み会🍺は楽しかったなぁ
⋮	⋮

表 3: 絵文字🍺の各意味内容での周辺語の TF-IDF 値

	ビール	飲む	忘年会	飲み会
飲む	0.0360	0.1001	0.0091	0.0254
ビール	0.1109	0.0041	0.0015	0.0016
する	0.0261	0.0269	0.0334	0.0222
忘年会	0.0017	0.0041	0.0927	0.0079
飲み会	0.0021	0.0015	*0.0000*	0.0777
今日	0.0100	0.0099	0.0198	0.0222
なる	0.0127	0.0111	0.0106	0.0111
てる	0.0067	0.0123	0.0137	0.0127
楽しい	0.0050	0.0082	0.0167	0.0111
会	*0.0000*	0.0020	*0.0000*	0.0323
行く	0.0044	0.0135	0.0046	0.0095
明日	0.0039	0.0064	0.0137	0.0063
会す	*0.0000*	*0.0000*	*0.0000*	0.0189
着物	*0.0000*	*0.0000*	*0.0000*	0.0152
プチ	*0.0000*	*0.0000*	0.0145	*0.0000*

示す．表 4 に示したように各意味のツイート量にはばら

表 4: コーパスの規模

	ツイート数	周辺語総数
ビール	188	1,936
飲む	164	1,821
忘年会	58	686
飲み会	48	678

つきがあるが，表 3 のようにそれぞれの特徴を TF-IDF 法で推定することが可能である．

3.2 SVM の学習と絵文字の意味分類

絵文字の意味による周辺語の出現特徴を利用して意味推定を行うため，サポートベクターマシン (SVM) を用いる．SVM とは，高次元特徴空間において線形関数の仮設空間を用いる教師あり学習システムである [3]．SVM の学習データとなる素性は，収集したすべてツイートに出現する単語それぞれの各ツイート中の出現数のベクトルとした．

作成したベクトルを用いて SVM で学習を行い分類を行う．SVM は二値分類器のため，正事例と負事例に分ける必要がある．例えば「ビール」に関する学習を行うときは，教師データ全体 1,048 ツイート中の「ビール」の意味で絵文字が使われている 188 ツイートを正事例，残りを負事例として学習する．学習した SVM を用いて，絵文字を含む入力ツイートがどの意味に属しているかを判別することによって，絵文字の意味推定を行う．

4 実験

絵文字「🍺」が含まれる 1,048 ツイートで実験を行った．評価方法として交差検証を行った．まず，教師データを 10 個に分割する．そのうちの 9 個の教師データを用いて学習を行い，残りの 1 つの教師データを用いて意味推定を行い，結果と実際の答えとの比較をする．これをすべての組合せで行なった．

提案手法による意味推定処理の結果を表 5 に示す．適

表 5: 交差検証の結果

	適合率	再現率	F 値
ビール	82.1%	83.5%	0.828
飲む	63.7%	66.4%	0.650
忘年会	70.1%	81.0%	0.752
飲み会	79.1%	79.1%	0.791

合率が最大で 82.1%，再現率が最大で 83.5%，F 値が最大で 0.828 と高い結果が得られた．この結果は，意味同士が似通っていて共起する単語に差が出にくいと想定されるにもかかわらず高い結果と言え，絵文字と共起する単語を元に意味を推定することができたと言える．「飲む」の各数値が他と比較して低い．これは，絵文字の意味が「飲む」以外の場合でも「飲む」という単語がよく使われており，「飲む」という意味の絵文字との共起語が，「飲む」以外の意味の絵文字との共起語と差が出にくく，他の意味として誤認したと考えられる．

5 まとめ

絵文字の用法に着目し，絵文字との共起語を元に絵文字の意味推定を行う手法を提案した．適合率，再現率，F 値で高い数値を得られ，用法に着目した絵文字の意味推定ができた．

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 回年次大会発表論文集, pp.567–570, 2010.

[2] 友山 孝広, 延澤 志保, “Twitter を対象とした絵文字の用例の自動分類,” 電子情報通信学会 2017 年総合大会学生ポスターセッション, No.ISS-P-39, p.39, 2017.

[3] Nello Cristianini, John Shawe-Taylor, 大北 剛, “サポートベクターマシン入門,”共立出版, 2005.