社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－NL－184 (11)
2008／3／28

日英対訳感情表現コーパスに基づく感情表現抽出手法の提案

松本 和幸 ・湊 純子 ・土屋 誠司 ・任 福継
徳島大学

本稿では，日英両言語における感情表現の特徴をコーパスに基づく統計的調査により導出し，得られた結果を用いた感情表現抽出手法の提案を行なう．感情タグ付きパラレルコーパスから得た感情カテゴリごとの感情表現の特徴には感情表現そのものが持つ特徴と周辺の形態素等が持つ特徴とがある．提案手法では，これらの複数の特徴を組み合わせることで，感情表現が示す感情の種類を判定する．具体的には，感情表現を構成する品詞，感情表現の出現位置，感情表現の前後の単語の品詞に着目し，感情表現の抽出を行なった．評価実験の結果，品詞特徴を用いた手法は“喜び”と“嫌悪”において判定精度90%以上という結果を得た．

キーワード：感性情報処理，感情タグ付きパラレルコーパス

Proposal of a method for extracting emotional expressions based on an emotion annotated Japanese-English parallel corpus

Kazuyuki Matsumoto , Junko Minato , Seiji Tsuchiya and Fuji Ren
The University of Tokushima

This paper statistically studies the emotional features of Japanese and English based on an emotion annotated parallel corpus and proposes a method for extracting emotional expressions. The proposed method estimates the emotion category of the emotional expressions by focusing on the three kinds of features: part of speech of emotional expression, position of emotional expression and part of speech of the previous/next morpheme of the target emotional expression. The evaluation experiment resulted over 90.0% (joy, hate) of accuracy in the method based on part of speech features.

Key Words: Sensibility information processing, Emotion tagged parallel corpus

1 はじめに

近年のインターネットの急激な普及により，外国語で記述されたテキスト情報にアクセスする機会が増えた．さらに，コンピュータを介したテキストベースのコミュニケーションの形態が我々の生活に浸透しつつある．このような状況において，Webなどを通じて日常的に世界各国で起きた事件について現地のメディアの報じ方，世論などをリアルタイムで知ることができるようになった．しかし，言語が異なるために，伝えられている内容を正確に理解できているとは言い難い．実際には，事件の内容についての適切な翻訳済みの記事を読むことが可能であるが，現地メディアの捉え方や世論については一部のみしか知ることができないといった問題がある．一方，Web上の記事における書き手の意図や態度，世論を知りたいという要求が高まっている．

そこで，ニュース記事から書き手の意見（態度）を抽出しようとする試みが行なわれてきた．四宮 [1] らの研究では，新聞記事中の単語に着目し，書き手の報道対象に肯定的な立場をとるか，

または否定的な立場をとるかを認識する手法を提案している．しかし，感情表現を収集して構築した辞書を用いる感情推定の場合，辞書に未登録の単語には対応できないといった問題が起こる．

我々が普段の会話において用いる感情表現は様々である．表1に示すように，表現形式や，各表現ごとに表す感情の程度も様々であるため，どのような表現をどういった状況において用いるかを判断することは，母国語であっても難しいといえる．

表1　感情表現の例

タイプ	例
単語	ぶんぶん，悲壮，がっくり，etc.
イディオム	腹が立つ，我を忘れる，etc.
四文字熟語	大驚失色，疑心暗鬼，etc.

「感情表現辞典」[4]は，現代小説で用いられている感情表現を10種類の感情に分類し，小説での実際の用いられ方を例文を踏まえて紹介している．これらのような辞典を用いることは日頃行なう文書作成などにおいては有効であると考えられるが，これらを用いて膨大なテキストリソースから表現を抽出するのは容易ではない．その理由として，感情表現の多様性が挙げられる．感情表現は単語で表現されるもの，句で表現されるものなど，様々な形式があり，一般的に使用される表現のみならず，日々創り出されるものである．

中山[5]らは，このような多様な表現に対処するため，係り受け関係を用いることで感情表現の自動抽出を試みている．

感情表現を用いることで，テキストベースのコミュニケーションが円滑に行なえることもあるが，一般に，母国語以外の言語で記述された感情表現を正確に理解するのは非常に困難であると考えられる．また，現在の機械翻訳技術の精度は十分ではなく，他国語により記述された感情表現を適切に翻訳することができないため，誤解を招く恐れもある．これらの問題を解決するため，我々は，様々な言語による文から感情表現を抽出し，抽出した表現から書き手の感情的態度を適切に示し，多言語コミュニケーションを支援するシステムの構築を目指している．このシステムは，外国語で記述された感情を含む記事の理解に役立てることができると考えられる．

本稿では，感情表現の用例文に基づいた日英パラレルコーパスに含まれる感情表現を分析することにより，感情表現が表す各感情毎の特徴を抽出する．そして，抽出された特徴を用いることで，感情表現が表現する感情を判定する手法を提案する．

2　日英対訳感情表現コーパス

感情表現を抽出するためには，単語がある感情をどの程度表現しているかを判定する必要がある．我々は，単語や文が感情を表現する程度を，“感情属性値”として定義している[6]．何らかの指標に基づき，感情表現の“感情属性値”を得ることで，文中に含まれる感情表現の抽出ができると考えている．

そこで，我々は，感情表現が日英の発話文においてどのように用いられるかを調査するため，対訳コーパスの構築を行なった．

「日英対照感情表現辞典」[7]から抽出した感情表現を含む対訳文1,190組に対し，日本語は各文に対してChaSen[8]を用いて形態素解析を行い，英語は各文に対してBrill's Tagger[9]を用いて品詞解析を行なった．そして，形態素解析結果に対して，1名の作業者により発話者の感情を示す文感情タグ（文単位）と，文中の単語やイディオムの感情を表す感情表現タグ（形態素単位）とを付与し，日英対訳感情表現コーパスの構築を行なった．

本研究では，基本となる感情の種類を表2に示す9種類とした．

表2　基本感情と下位感情

基本感情	下位感情
怒り	不平不満, 非難
不安	恐れ, 苦しみ, 狂った
喜び	期待
嫌悪	軽視・軽蔑
愛	羨望・願望，哀れみ・共感，執着
尊敬	賞賛
悲しみ	後悔, 恥, 惨めさ, 孤独, 絶望
驚き	興奮
平静	-

以下，日英対訳感情表現コーパスに付与したタグについて説明する．

2.1　文感情タグ

文感情タグは，発話文の発話者（書き手）の感情の種類を表すタグであり，1文につき1〜3種類までを付与している．話者感情として上位感情に適当なものがなければ下位感情から選択し，上位感情を文感情タグとして付与した．

2.2 感情表現タグ

感情表現タグは，感情表現が1つの形態素のみから構成される場合と複数形態素から構成される場合とを区別するために付与する“感情表現タイプタグ”と，感情表現が表す感情の種類を表す“感情属性タグ”とに分類される．感情表現タイプタグを表3に示す．また，感情属性タグを表4に示す．

表3 感情表現タイプタグ

タグ	意味
S	感情語
B	感情イディオムにおける最初の形態素
I	感情イディオムにおける中間の形態素
E	感情イディオムにおける終端の形態素

表4 感情属性タグ

ID	タグ	感情の種類
1	anger	怒り
2	anxiety	不安
3	hate	嫌悪
4	joy	喜び
5	love	愛
6	respect	尊敬
7	sorrow	悲しみ
8	surprise	驚き
9	neutral	平静

2.3 基本統計結果

構築したコーパスの基本統計結果を表5に示す．

本コーパスは1文につき感情表現が少なくとも1つ以上含まれる文のみで作成した．しかし，文中の感情表現が，発話文の話者（書き手）の感情を表現しているとは限らないため，“話者感情を付与された文の総数”は総文数と等しくはならなかった．

また，“neutral”以外の感情属性タグが付与された表現であっても，常に同じ感情属性が付与されているわけではなかった．さらに，“neutral”が付与された表現にも“neutral”以外の感情属性タグが付与される場合もあった．これは，感情表現の用い方によっては感情を表現しない場合があるためである．このような例として，以下のようなものがある．

- He is very kind to his wife.

  “kind”には，感情属性タグ“love”が付与されている．

- I hate such a kind of impudent fellow as he.

  “kind”は感情属性タグが付与されていない．

3 感情表現の特徴分析

本稿では，感情表現の特徴に基づいた感情判定を行うため，次の3つの観点に基づき，特徴分析を行った．

(A) 感情表現を構成する品詞分布
(B) 感情表現の文中での出現位置分布
(C) 感情表現の前後に出現する品詞の組み合わせ分布

以下，上記の感情表現の特徴の分析方法について述べる．

3.1 感情表現の品詞分布

まず，(A)については，感情表現がどのような品詞であるかを調査した．品詞体系は，日本語はIPA品詞体系，英語はBrill's Taggerの採用する品詞体系に則した．日本語の品詞分類の例を表6に示す．

表6 品詞分類（日本語: IPA品詞体系）

大分類	小分類
名詞	名詞-一般, 名詞-固有名詞, etc.
代名詞	名詞-代名詞-一般, etc.
接頭詞	接頭詞, 接頭詞-名詞接続, etc.
動詞	動詞, 動詞-自立, 動詞-非自立, etc.
形容詞	形容詞, 形容詞-自立, etc.
副詞	副詞 , 副詞-一般, etc.

日英における感情表現を構成する形態素の品詞分布を表7，8に示す．これら2つの統計結果を日英で比較すると，次のような特徴を確認できた．

表 5　コーパスの基本統計結果

種類	日本語	英語
総文数	1190	1190
総単語数 (Unique)	14202(2131)	11235(2409)
1 文における単語数平均	11.93	9.44
話者感情を付与された文の総数	601	601
感情語の総数 (Unique)	1220(610)	1249(894)
感情イディオムの総数 (Unique)	274(231)	259(248)
修飾語の総数 (Unique)	108(70)	39(35)
否定語の総数 (Unique)	88(26)	31(15)
1 文における感情表現数の平均	1.26	1.27
感情表現に感情属性タグ 'neutral' が付与されている割合 (%)	0.22	0.15

- 日英両言語とも，“形容詞 (Adjective)”, “動詞 (Verb)” で感情表現されることが多い
- 日英両言語とも，“副詞 (Adverb)” で感情表現されることは少ない

本研究において構築した感情コーパスは小規模なため，これらの特徴は本コーパスに特有の特徴である可能性があるが，感情表現に関して日英に共通するものと異なるものを見極める際の参考になると考えられる．

表 7　感情属性ごとの品詞分布（日本語）

	1	2	3	4	5	6	7	8
動詞	124	109	132	143	136	31	119	18
副詞	6	19	12	29	6	5	15	6
形容詞	12	18	205	81	57	57	68	1
名詞	53	39	92	61	42	18	59	4
感動詞	0	0	2	3	0	1	0	0
助詞	22	24	30	22	33	7	34	2

3.2　感情表現出現位置の分布

日英両言語における感情表現の出現位置の分布を調査した．

まず，文中に含まれる感情表現の出現位置 ($EApos$) を式 1 により計算する．式 1 中の単語の位置 ID とは，文頭からの単語の通し番号で，1 から始まる整数で表す．

表 8　感情属性ごとの品詞分布（英語）

	1	2	3	4	5	6	7	8
動詞	90	67	106	113	99	26	76	14
副詞	4	2	6	6	3	3	8	1
形容詞	36	52	209	85	36	59	75	4
名詞	60	60	92	103	76	29	80	6
感動詞	0	0	0	0	0	1	0	0
代名詞	1	3	0	4	11	2	4	0
限定詞	4	5	4	6	6	1	3	1

$$EApos = \frac{\text{単語の位置 ID}}{\text{文中の単語総数}} \tag{1}$$

また，計算した $EApos$ は，表 9 に示す区間毎に分けて表現する．

3.3　感情表現の前後に出現する語の品詞分布

感情表現の前後に出現する品詞を特徴として得る．具体的には，感情表現 w_i の前後の品詞を $POS(w_{i-1})$, $POS(w_{i+1})$ で表すと，その組み合わせと，感情表現に付与されている感情属性タグの共起頻度を計算する．この共起頻度を正規化した値を，前後の語の品詞と感情属性との関連性を示すものとして用いる．

表 9　感情語・イディオム出現区間

区間番号	$EApos$ の範囲
0	$0.00 \leq EApos < 0.05$
1	$0.10 \leq EApos < 0.15$
2	$0.15 \leq EApos < 0.25$
3	$0.25 \leq EApos < 0.35$
4	$0.35 \leq EApos < 0.45$
5	$0.45 \leq EApos < 0.55$
6	$0.55 \leq EApos < 0.65$
7	$0.65 \leq EApos < 0.75$
8	$0.75 \leq EApos < 0.85$
9	$0.85 \leq EApos < 0.95$
10	$0.95 \leq EApos \leq 1.00$

4　評価実験

本稿では，まず，文中から感情表現を抽出するために，感情表現が感情を持つか持たないかのみに限定した判定を行なう．

文中に含まれる語に付与されている感情表現タグが “neutral” かを判定するために，まず，抽出には感情極性対応表 [10] を用いる．感情極性対応表には，単語とその感情極性が登録されており，日英両言語のものが公開されている．単語が感情極性対応表中に含まれており，感情極性値が閾値を超えればその単語を感情表現とする．含まれていないか，閾値を超えない場合は “neutral” と判定する．

今回，閾値を 0.5 として判定実験を行なった．感情極性対応表にはイディオムは含まれていないため，単語のみを抽出対象とした．実験の結果，感情属性タグの付与されたタグの抽出精度は，日本語では 38.9%(474/1220)，英語では 46.8%(585/1249) となった．

次に，この実験で抽出に成功した感情表現を対象とした感情表現の感情判定実験を行なう．

4.1　実験-(1)：品詞特徴を用いた単語感情判定

統計分析の結果，感情表現は品詞毎の感情属性タグの偏りが見られた．ここでは，品詞毎の感情表現出現頻度に基づき，単語感情判定を行なう．

単語感情判定に用いる特徴抽出の流れを，以下に示す．

(1) まず，コーパス中に含まれる感情表現の感情属性タグとその品詞との組み合わせをすべて抽出する．
(2) 抽出した組み合わせに基づき，品詞を大分類に変換し，品詞の大分類毎の感情属性タグの付与頻度を算出する．
(3) 付与頻度を正規化し，品詞毎の感情属性重みを算出する．

得られた感情属性ごとの品詞特徴を用いて，入力された感情表現の感情判定実験を行なった．実験は交差検定 (*Leave-one-out Cross Validation* 法) で行なった．実験対象は，抽出に成功した表現 (日本語 474, 英語 585) とした．評価は，コーパスに付与されている単語感情タグ (正解タグ) と，推定結果である感情属性タグとの一致の回数に基づく一致率を用いて行った．式 2 に一致率 MR の計算式を示す．$MatchCount(Est, Ans)$ は，推定感情属性タグと一致した，正解タグの数を表す．また，$Count(Est)$ は，推定された感情属性タグの総数を表す．また，今回，推定感情属性タグは上位 2 位までを推定結果とした．

$$MR = \frac{MatchCount(Est, Ans)}{Count(Est)} \tag{2}$$

実験結果を表 10 に示す．

表 10　実験結果-(1) の結果

感情属性	一致率	
	日	英
anger	0.0%	21.6%
anxiety	26.3%	2.1%
hate	96.6%	98.8%
joy	100.0%	100.0%
love	0.0%	0.0%
respect	0.0%	3.5%
sorrow	63.0%	94.0%
surprise	0.0%	0.0%

実験の結果，“喜び”，“嫌悪”，“悲しみ” の一致率が比較的高く，その他の感情属性に関しては低い一致率となった．

4.2　実験-(2)：出現位置特徴を用いた単語感情判定

感情表現の出現位置を分析したところ，日英ともに，感情毎の分布に偏りがあった．そこで，感情表現の出現位置毎の頻度を用いることで単

語感情判定を行なう．

実験-(1) と同様，感情表現（感情語・イディオム）をどの程度正確に感情判定可能であるかを評価するため，交差検定による感情表現の感情判定実験を行なった．交差検定には，実験-(1) と同様，*Leave-one-out Cross Validation* 法を用いた．実験対象は，実験-(1) と同じものを用いた．また，評価方法も実験-(1) と同じにした．実験手順は以下の様になる．

(1) まず，式 1 を用いて各感情表現の出現位置の区間番号を取得する．

(2) 出現位置区間毎の感情表現タグの付与頻度 $(freq(i, E_x))$ から，区間毎の感情重みを式 3 により計算する．式中のi,mは出現位置の区間番号，x,nは感情属性タグの ID を示す．

この式によりコーパス中の感情表現の出現位置から，感情表現の表す感情を決定する．

$$EA_{x,i} = \frac{freq(i, E_x)}{\sum_{m=0}^{10} \sum_{n=1}^{9} freq(m, E_n)} \quad (3)$$

実験-(2) の結果を，表 11 に示す．

表 11　実験-(2) の結果（一致率）

日	英
44.6%	59.3%

4.3 実験-(3): 前後の品詞特徴を用いた単語感情判定

感情表現の前後に出現する語の品詞の組み合わせを特徴として，感情表現の感情判定を行なった．前後に出現する品詞は大分類に変換した上で，コーパスから特徴抽出を行った．実験対象は実験-(1) と同じものを用いた．また，評価方法は実験-(1) と同様，一致率を用いた．評価実験結果を表 12 に示す．

表 12　実験-(3) の結果（一致率）

感情属性	一致率	
	日	英
anger	45.9%	39.2%
anxiety	57.9%	36.2%
hate	83.8%	67.7%
joy	78.0%	65.8%
love	36.4%	47.6%
respect	25.0%	63.2%
sorrow	41.2%	42.9%
surprise	0.0%	0.0%

5 まとめと今後の課題

本稿では，日英両言語における感情表現の特徴分析を，我々が構築した感情タグつきのパラレルコーパス（日英対訳感情表現コーパス）を用いて行なった．そして，感情表現抽出および感情判定手法の提案を行なった．提案手法は単純な特徴のみを用いているが，評価実験を行なった結果，感情の種類によっては高い判定精度を得ることができた．

本手法は感情表現がどの感情を最も強く表しているかの判定をコーパスから抽出した特徴に基づき行なうことを目的としているが，今回用いた抽出手法はコーパスに基づいた手法ではなく，抽出精度があまり高くなかった．

また，感情表現であるか否かを判定する場合に，コーパスからの特徴量を用いる方法も考えられるが，感情表現以外の語彙は様々であり，現在の規模のコーパスから特徴量を抽出したのみでは十分な精度が得られないことが予想される．

今後は，本稿で用いた品詞や出現位置特徴と，単語の持つ意味的な情報を組み合わせた感情表現抽出・判定手法を提案したいと考えている．

謝　辞

本研究の一部は，科学研究費基盤研究 B（課題番号 19300029）の助成を受けて行われた．

参考文献

[1] 四宮 瑞穂, 三品 賢一, 土屋 誠司, 任 福継: “新聞記事の意見抽出のための感情語辞書の有効性に関する考察”, 信学技報, Vol. 107, No. 387, TL2007-47, pp.37–42, 2007.

[2] 遠藤大介, 齋藤真実, 山本和英: “係り受け関係を利用した感情生起表現の抽出”, 言語処理学会第 12 回年次大会講演論文集, pp.947–950, 2006.

[3] 佐藤 一誠, 平手 勇宇, 山名 早人: “距離と属性を考慮した PrefixSpan による感情表現抽出”, 電子情報通信学会第 17 回データ工学ワークショップ第 4 回日本データベース学会年次大会 (DEWS-2006), 2006.

[4] 中村明, “感情表現辞典”, 東京堂出版, (1993).

[5] 中山 記男, 江口 浩二, 神門 典子: “感情表現の抽出手法に関する提案”, 情報処理学会研究報告 2004-NL-164, pp.13-18, 2004.

[6] 松本 和幸，三品 賢一, 任 福継, 黒岩 眞吾: “感情生起事象文型パターンに基づいた会話文からの感情推定手法”，自然言語処理, 言語処理学会, Vol. 14, No. 3, pp.239–271, 2007.
[7] 稗島 一郎: “日英対照 感情表現辞典”， 東京堂出版，1995.
[8] “日本語形態素解析システム ChaSen 「茶筌」”. http://chasen.naist.jp/hiki/ChaSen/
[9] Brill, E. , “A Simple Rule-based Part-of-Speech Tagger.” *In Proceedings of 3rd Applied Natural Language Processing*, pp.152–155, 1992.
[10] 高村 大也，乾 孝司，奥村 学: “スピンモデルによる単語の感情極性抽出”, 情報処理学会論文誌, Vol.47, No.2 pp.627–637,2006.