感情推定における若者言葉の影響

松本 和幸　　任 福継
徳島大学大学院ソシオテクノサイエンス研究部
{matumoto, ren}@is.tokushima-u.ac.jp

1　はじめに

近年，人間と自然言語を用いてコミュニケーションのできるコンピュータに応用するための感情推定技術の研究が盛んである。その中でも，テキスト情報からの感情推定に関する研究は，評判・意見分析やコミュニケーションロボットへの応用などが期待され，注目されている [1]．我々は，自然言語文からの感情抽出・感情推定技術に関する研究を進めている [2, 3, 4]．文から書き手の感情推定を行うために，文中に含まれる単語や句がどのような感情を表すかを単語辞書に登録しておき，それを用いる方法が考えられる．しかし，多くの言語表現と同様に感情表現は無数にあり，単語辞書に基づいて感情推定できる文は限られている．しかも，Web 上の膨大な言語情報には大量の未知の表現が含まれ，従来では例外とされてきたような文は，これからも増えていくと考えられる．このことから，既存の辞書に登録されていないような感情表現を含む文から感情推定を行う手法が必要になることが予想される．

Weblog などにおける文書には，形態素解析器で正しく解析できない表現が多く含まれている．それらの多くは，未知語として判定されるか，誤分割されてしまう．その中でも，若者が日常生活や Web 上で用いる表現（若者言葉と呼ばれる）は，非常に多くのバリエーションがあり，日々増加している．これら若者言葉は，既知語を面白おかしく言い換えたものや，既知語にはない意味を持つ語など，様々なタイプがあるため，すべてをまとめて，若者言葉と一括りとすることは難しく，明確に定義することは不可能に近い [5, 6]．

従来，自然言語処理の分野では，辞書に登録されていない未知語の解析については盛んに研究が進められてきた．しかし，対象となる未知語の多くが固有名詞やオノマトペ，顔文字などであり，俗語，特に若者言葉を主な対象とした研究は少なかった．

若者言葉の中には，その語が出現した当初は若者が多く使用していたが，時間が経過するにつれ，一般的に用いられるようになった語も存在する．たとえば，「やばい」という語は，元々若者言葉であったが，現在では一般的に用いられるため，形態素解析用の辞書にも形容詞として登録されている．このことから，若者言葉といえども一過性のものではなく，将来的に使い続けられる可能性もあるため，軽視することはできない．

久保村 [7] らは，若者言葉をタイプ別に分類し，省略型の若者言葉を，辞書に登録されている語（既知語）に変換する手法を提案している．しかし，実際には，既知語に変換できるような省略型の語は限られている．また，感情を表現するような若者言葉は，既知語に変換してしまうと，微妙なニュアンスが変化してしまうため，変換後に感情推定を行う方法は，あまり有効ではないと考えられる．さらに，若者言葉は，既知語に置き換えることが不可能な新しい概念の語や，既知語と同じ表記であるが若者言葉としてはまったく別の意味になる語も存在する．こういった語が感情を表現する場合，既知語に変換する方法を用いても効果が期待できない．そこで本研究では，若者言葉による感情表現の特徴を用例から調査する必要があると考え，若者言葉を含んだ文を収集することで，コーパスを構築し，分析を行う．そして，収集した文を用いた感情推定実験を行い，考察する．

2　若者言葉感情コーパス

我々は，若者言葉を含んだ文ばかりを集め，それらの文に書き手の感情を表す感情タグおよび，含まれている若者言葉を付与することで，若者言葉感情コーパスを構築した．実際に Weblog 等で使用されている例文を集めるため，Yahoo! Blog 検索 [8] を用いることで，若者言葉を含む文の自動収集を試みた．自動収集した文には，タイトルや重複文などのノイズが含まれているため，手作業により取捨選択を行った．付与する感情タグは，“喜び”, “怒り”, “嫌悪”, “期待”, “不安”, “受容”, “尊敬”, “愛”, “驚き”, “後悔”, “悲しみ”, “恥”, “平静” の 13 種類であり，1 文につき，複数種類のタグを付与することを許可する．作業者 2 名で構築を行い，現時点で 9014 文に対してタグ付けが完了している．表 1 に，若者言葉感情コーパスの例を示す．このデータをもとに，若者言葉毎の特徴の分析を行う．

3　若者言葉毎の特徴分析

若者言葉感情コーパスに登録されている文に含まれている若者言葉の種類の異なり数は，表記違いを除くと，全部で 1187 種類である．若者言葉感情コーパスにおける，各感情タグの付与の割合を，表 2 に示す．また，コーパス中の出現頻度が 100 以上の若者言葉 (6 種類) について，それらの語が出現した文に付与された各感情タグの付与割合を統計してみた．その結果を表 3 に示す．

若者言葉同士の共起頻度を計算してみたところ，最も多くの若者言葉と共起している若者言葉は，“マジ”(70

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

;

表 1: 若者言葉感情コーパスの例

若者言葉	感情タグ	文
やばい	不安	むしろ悪化してるやばい。
ぶっちゃけ, うざい	嫌悪	ぶっちゃけ … 少々うざくなっている今日この頃 …
超, むかつく	怒り	超むかつくんだけど。

表 3: 出現頻度 100 以上の若者言葉毎の感情タグの付与割合 (%)

	怒り	喜び	悲しみ	不安	驚き	嫌悪	期待	尊敬	愛	恥	後悔	受容	平静	頻度
マジ	**25.0**	10.2	2.6	7.1	0.5	5.1	22.4	11.7	7.1	4.6	0.0	0.0	3.6	196
めっちゃ	2.5	**23.2**	5.8	10.4	0.4	5.8	7.9	12.0	7.1	7.9	0.4	0.0	16.6	241
やばい	3.0	10.1	3.0	**45.0**	3.0	9.5	13.0	3.6	2.4	2.4	0.0	0.0	5.3	169
きもい	6.5	3.6	1.4	8.0	0.0	0.0	**76.1**	2.9	0.0	0.7	0.0	0.0	0.7	138
うざい	40.7	1.6	2.4	4.9	0.0	0.0	**43.9**	1.6	0.0	0.0	0.8	0.0	4.1	123
むかつく	**74.4**	1.0	1.5	0.0	0.0	0.5	13.3	2.0	0.5	0.5	0.0	1.0	5.4	203

表 2: 感情タグの付与割合

感情極性	感情タグ	付与数	割合 (%)
positive	期待	1672	17.76
	喜び	1557	16.54
	受容	1065	11.31
	愛	256	2.72
	尊敬	143	1.52
neutral	驚き	285	3.03
	平静	203	2.16
negative	嫌悪	2333	24.78
	怒り	796	8.46
	不安	728	7.73
	悲しみ	309	3.28
	恥	34	0.36
	後悔	33	0.35

回)，その次が “超”(56 回) であった．“マジ” や “超” は，他の語，主に形容詞の接頭語として用いられる語である．これらの接頭語は，それのみでは感情を表現することはできないが，形容詞となる語を修飾することで，その語の意味を強めたり弱めたりすることができる．そのため，感情の強さに影響すると考えられる．また，同様に，単独での感情表現には適さない “ぶっちゃけ” や，“やっぱ” なども，他の若者言葉との共起回数が多かった．また，顔文字は，書き手の感情状態を表現するために用いられることが多い [9]．顔文字を含んでいる文は，コーパス中，496 文であり，全体の約 5.5%程度であった．

4 感情推定手法

若者言葉を含む文は，新聞記事などの文と比べると語順が文法に則っていなかったり，助詞が省略されたり，語尾がくだけた表現で書かれやすい性質がある．こういった文の場合，前処理である形態素解析や構文解析の精度が低くなることが予想される．

江村 [10] らは，Mobile Weblog(モブログ) を対象とした書き手の感情推定を行う際に，サポートベクトルマシン (SVM) を用い，学習素性としては，ストリングカーネルを採用している．ストリングカーネルを用いることで，形態素解析時の誤解析の影響は抑えられる．しかし，これまで様々なテキスト分類の研究において，形態素は意味のある最小の単位であるため，素性として用いられることが多かった．特に，感情推定のようなタスクにおいては，単語の意味は非常に重要であるため，形態素を素性とすべきである．小川 [11] らは，Web 上の比較的短い文における感情推定を行うため，形態素 1-gram を用いた．また，山本 [12] らは，感情コーパスを半自動的に構築するため，文への感情タグ付与を，ナイーブベイズ分類器を用いて行った．また，三品 [13] らの研究では，形態素 N-gram に基づく類似度を用いた独自の感情推定手法の改良を行った結果，4 種類の感情推定で約 80%の推定精度を得ている．また，若者言葉の感情コーパスを扱った従来研究として，松本 [14] らがナイーブベイズ分類器と蓄積手法を用いて感情推定実験を行っている．

しかし，これらの研究では，若者言葉のような文中の未知語が与える感情推定への影響については詳しく分析されておらず，未知語に対して有効な手法であるかどうかを判断できない．若者言葉を含む文に適した感情推定手法を提案するためには，これらの研究で用いられた手法が，未知語を含み，正しい解析が困難な文に対して有効かどうかを調べる必要がある．そこで，本研究では，文中の若者言葉がどの程度感情推定に影響するかを確認するため，まず，単純な素性として形態素の表層形を用いることとする．さらに，未知語と認識される若者言葉および既知語と認識される若者言葉，また，感情を表現する既知語や，顔文字なども素性として用いることにする．次に，文からの素性抽出および分類実験の流れを示す．

1. まず，文を MeCab[16] を用いて形態素解析し，感情表現として用いられやすいと考えられる語（感情語 EF）を，形容詞や副詞などを登録した感情語辞書 [15] を参照して抽出する．また，感情語に付与されている感情タグ ETF も素性として考慮する．

2. 感情語とは別に，顔文字辞書を用いて顔文字と判断できる文字列 (FF) を抽出する．用いた顔文字

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

辞書は，Web 上で公開されている顔文字辞書 7 種類から顔文字の顔の部分（括弧で囲まれている部分）のみを抽出したものである (登録数：8533).

3. 抽出した感情語および顔文字を素性として SVM またはナイーブベイズ分類器 (NB) を用いて学習させ，分類実験を行う．本研究では，サポートベクトルマシンとして，多クラス分類が行える SVM ^{multiclass} [17] を用いた．SVM ^{multiclass}での学習においては，デフォルトの線形カーネルを用いてパラメータ c の値は 0.3 とした．

5 感情推定実験

本実験では，感情推定において，若者言葉 (WF) を素性とした場合と，若者言葉以外の感情表現も素性とした場合とで，感情推定の精度を比較する．また，ベースラインとして，文中から抽出した形態素 N-gram を素性とした実験を，1-gram から 5-gram まで，それぞれ行った．

5.1 実験概要

感情推定結果の評価方法として，SVM または NB により推定された感情の種類が，コーパスに付与されている感情タグの種類と一致していれば正解，一致していなければ不正解とし，正解率を出す．推定する感情の種類は，感情タグの種類と一致する 13 種類と，これらの感情を positive/negative/neutral に分けた 3 種類とした．

実験に用いる素性の組合せを，表 4 に示す．

表 4: 素性の組合せ

1	WF
2	$WF + EF$
3	$WF + EF + FF$
4	$WF + ETF + FF$

評価方法として，10 分割交差検定を用いた．対象となるデータをランダムに 10 等分し，1/10 をテストデータ，残りの 9/10 を学習データとする．これを，テストデータを入れ替えて 10 回繰り返し，精度の平均を計算する．本実験で用いた文は，構築したコーパス中のすべての文 (9014 文) である．

5.2 実験結果

感情推定実験の結果を，表 5 に示す．13 種類の感情推定で，素性 $EF + FF + WF$ を用いた時の NB による実験結果を若者言葉毎に統計した結果，出現頻度 50 以上の若者言葉で，推定精度 60%以上のものと，推定精度 40%以下のものの一部を表 6 に示す．

また，ベースラインの形態素 2-gram 素性を用いた SVM による実験結果から，未知語含有率と推定失敗率について，分析を行った．図 1 に，未知語含有率と推定失敗率の関係を示す．図中の横軸は，各文における未知語含有率 (%) を表し，縦軸は感情推定失敗率 (%) を表す．これより，文の未知語含有率の高さが感情推定失敗の主な原因とまではいえないことが分かる．

表 5: 感情推定実験結果（正解率 %）

素性	13 種類		3 種類	
	SVM	NB	SVM	NB
WF	37.70	39.38	65.14	67.97
$EF + WF$	38.24	41.17	66.13	70.00
$EF + FF + WF$	38.14	41.67	66.13	69.96
$ETF + FF + WF$	35.59	41.50	64.25	70.09
1-gram	9.04	40.60	15.57	71.32
2-gram	31.77	37.85	51.80	68.00
3-gram	29.80	32.77	56.33	61.48
4-gram	23.19	28.84	54.54	56.26
5-gram	18.82	27.01	53.68	54.08

表 6: 若者言葉毎の感情推定正解率（%）

正解率 60%以上		正解率 40%以下（昇順）	
若者言葉	正解率	若者言葉	正解率
ありござ	98.0	ガングロ	15.4
かわええ	77.3	ヤン車	17.2
サイコー	76.9	ボコる	19.4
きもい	75.0	パソ禁	20.2
むかつく	73.8	イケメン	20.3
めどい	72.5	借りパク	20.4
引退フラグ	68.2	寝落ち	21.1
きしょい	67.1	アド変	23.4
パネェ	64.4	パチモン	24.2
プギャー	62.5	はぶる	25.0

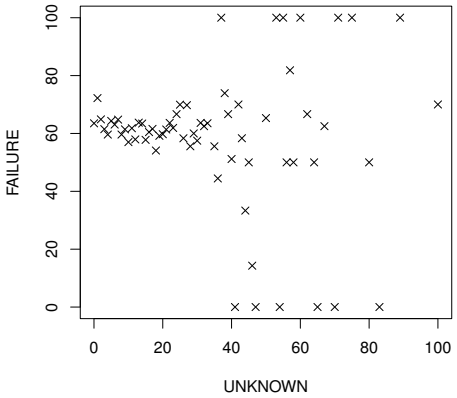

図 1: 未知語含有率と推定失敗率

5.3 考察

実験結果から，若者言葉を素性に含めた場合，SVM による 13 感情の推定では，ベースライン (形態素 N-gram を素性とする) を平均して 14.9%上回る精度が得られた．これは，形態素 N-gram を用いる場合よりも素性の種類数は圧倒的に少なく，その分，誤推定原因

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

となるノイズ素性が少なかったためであると考えられる．また，コーパスの各文は短い文が多いため，文中における各素性の出現頻度は，ほとんどが 1 となる．本実験では素性の値として出現頻度を用いたが，各素性の重要度を学習データから計算し，重要度の高いもののみを素性として用いることで，より推定精度を高めることができると考えられる．

また，文中に含まれる若者言葉毎に正解率を計算したところ，それのみで感情を表現できるような若者言葉が含まれる場合に，正解率も高くなるという結果が得られた．一方で，若者言葉それ自体のイメージや表現する感情に関わらず，同一文中で共起する感情表現や顔文字などが原因で，感情推定に失敗することもあるため，素性の組み合わせ方の改善も必要である．

6 おわりに

本稿では，テキストからの書き手の感情推定において形態素解析ではほとんどの場合，未知語として解析される若者言葉に着目し，若者言葉が含まれる文からの感情推定実験を行うことで，感情推定における若者言葉の影響について考察した．SVM，NB により，若者言葉，感情語，顔文字を素性として学習した実験の結果，平均的に，形態素 N-gram を用いたベースライン手法よりも高い推定精度が得られた．しかし，3 種類 (positive/negative/neutral) の感情推定に関しては，ベースラインの形態素 1-gram を素性とした NB を用いたときに，最大の推定精度 (71.32%) が得られた．このことから，positive/negative/neutral といった単純な種類の感情推定を行う場合には，形態素 1-gram は有効であることがいえる．一方で，13 種類の感情推定においては，若者言葉，感情語，顔文字を素性とすることで，ある程度の精度を得ることができることが分かった．positive/negative/neutral の推定後に，13 種類の感情推定を行うことで，推定精度をより高めることができると考えられる．

今後は，誤推定された文について，より詳細な分析を行い，感情推定において重要でない素性を求める方法を考えたい．

謝辞

本研究の一部は、科学研究費補助金（挑戦的萌芽研究：21650030）により実施した。

参考文献

[1] 大塚裕子, 乾孝司, 奥村学. 意見分析エンジン. コロナ社, 2007.

[2] F. Ren. From cloud computing to language engineering, affective computing and advanced intelligence. *International Journal of Advanced Intelligence*, Vol. 2, No. 1, pp. 1–14, 2010.

[3] K. Matsumoto and F. Ren. Estimation of word emotions based on part of speech and positional information. *Computers in Human Behavior*, 2011 in press.

[4] C. Quan and F. Ren. A blog emotion corpus for emotional expression analysis in chinese. *Computer Speech and Language*, Vol. 24, pp. 726–749, 2010.

[5] 米川明彦. 若者語を科学する. 明治書院, 1997.

[6] 北原保雄. あふれる新語. 大修館書店, 2009.

[7] 久保村千明, 原田俊信, 佐々木洋輔, 山本義人, 亀田弘之. ブログ記事を素材とする若者語処理システム評価方法. 信学技報, Vol. 105, No. 615, pp. 165–169, 2006.

[8] Yahoo! Blog 検索. http://blog-search.yahoo.co.jp/.

[9] 加藤尚吾, 加藤由樹, 小林まゆ, 柳沢昌義. 電子メールで使用される顔文字から解釈される感情の種類に関する分析. 日本教育情報学会学会誌, Vol. 22, No. 4, pp. 31–39, 2007.

[10] 江村恒一, 安木慎, 宮崎誠也, 久保山哲二, 青木輝勝, 安田浩. Svm を用いたモブログテキストからの感情抽出. 信学技報, Vol. 106, No. 473, pp. 61–66, 2007.

[11] 小川拓貴, 松本和幸, 任福継. “えもにゅ” における短文の感情推定について. 情処研報, Vol. 2010-NL-195, pp. 1–6, 2010.

[12] 山本麻由, 土屋誠司, 黒岩眞吾, 任福継. 感情コーパス構築のための文中の語に基く感情分類手法. 情処研報, Vol. 2007, No. 76, pp. 31–35, 2007.

[13] 三品賢一, 土屋誠司, 鈴木基之, 任福継. コーパスごとの類似度を考慮した用例に基づく感情推定手法の改善. 自然言語処理, Vol. 17, No. 4, pp. 91–110, 2010.

[14] K. Matsumoto, Y. Konishi, H. Sayama, and F. Ren. Wakamono kotoba emotion corpus and its application for emotion estimation. *Progress of Advanced Intelligence*, Vol. 2, pp. 89–99, 2010.

[15] 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, Vol. 14, No. 3, pp. 239–271, 2007.

[16] MeCab. : Yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.net/.

[17] T. Joachims. Svm multiclass multi-class support vector machine. http://svmlight.joachims.org/svm_multiclass.html.

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.