

係り受け関係を利用した感情生起表現の抽出

遠藤 大介, 齋藤 真実, 山本 和英

長岡技術科学大学 電気系

E-mail:{endo,saito,ykaz}@nlp.nagaokaut.ac.jp

1 はじめに

近年のインターネットの普及に伴い、電子メールやチャットなどで大量のテキストが個人間でやり取りされている。人間が実際対面して会話をする際には、発話した言葉以外に相手の表情や声の抑揚などの多くの情報から相手の感情を読み取ることができる。しかし、インターネットでのやり取りでは相手の書いた文章から感情を読み取る必要がある。そういった場面でコンピュータがユーザーの手助けをするシステムは有用性の高いものになる可能性がある。最近では、評判情報の抽出や電子メールの文章から時間表現や感情表現を抽出する研究 [1] が行われている。しかし、コンピュータにとって文章中から人間の感情を推定することは非常に難しく多くの課題がある。我々は、コンピュータが感情を推定するためには感情表現の辞書とともに感情が起こる原因となる表現、つまり感情生起表現の辞書が必要であると考え。ここで感情生起表現とは、感情が生起される要因の動作や事柄の表現のことである。例 1 に感情表現、例 2 に感情生起表現の例を示す。

例 1) うれしい、腹立たしい

例 2) 遊園地に行く、自分だけ処分される

本稿では Web 文書中の感情表現とその係り受け関係から書き手の感情が推定可能な表現を抽出する手法を提案する。

2 関連研究

田中ら [2] は情緒生起情報を結合価パターンに付与することによって、日記の情緒推定を行っている。また、中山ら [3] は感情表現の抽出に際し、「うれしい」と「面白い」を種とし prefixspan を用いて 2 語以上の感情表現のパターンを抽出し、そのパターンを利用し 1 語の感情表現を取ってくるというブートストラップを行っている。これらの研究は 2 語以上の表現を抽出する際にパターン辞書を用いているが、本稿ではパターン辞書を作成するのではなく、感情生起表現を抽出したものをフレーズとして保持する辞書の作成を目指している。

3 感情生起表現の定義

文書中の「うれしい」「かなしい」といった表現を感情表現とすると、感情生起表現とは「プレゼントをも

らう」のように書き手の感情が生起される要因となった事柄を指す。また、本稿では感情生起表現を文中で連続した 2 文節の表現と定義し抽出する。例 3 の文の場合、「プレゼントをもらえたことが」という表現を抽出する。ここで「彼に」の部分も感情の生起に関係しているがそこまで抽出した場合、表現の数が膨大になるため本稿では抽出文節を 2 文節としている。

例 3)

私は彼にプレゼントをもらえたことがうれしかった。

4 提案手法

本稿では以下の手順で感情生起表現の抽出を行った。

- (1) 種用感情表現辞書の作成
- (2) 生起表現候補の収集
- (3) 生起表現候補のフィルタリング

4.1 種用感情表現辞書の作成

感情表現を含む文を抽出する際に使用する種用感情表現辞書を以下の手順で作成する。感情表現辞典 [4] では感情を 10 分類 {喜、怒、哀、怖、恥、好、厭、昂、安、驚} に分類している。本稿でもこの 10 分類を採用した。まず、感情表現辞典の語句編に記載されている感情表現 10 分類全 2167 語を、それぞれの感情ごとに感情辞書として抜粋する。文の抽出に使用する種用感情表現辞書には、感情辞書から茶筌 (1) で定義されている「形容詞」または「名詞-形容動詞語幹」を含む表現であり、かつ「助詞-格助詞」を含まないものを抽出し登録する。例 4 に種用感情表現辞書として抜き出した感情表現の例を示す。また、種用感情表現辞書の感情ごとの感情表現数を表 1 に示す。

例 4) 種用感情表現辞書に登録されている語句例

喜：楽しい、にこやか、気楽、面白い、幸せ
怖：不気味、怖さ、心細い、不安、恐ろしい
安：平静、気安い、気楽、和やか、悠長

例にあげた「気楽」のように二つ以上の感情に登録されている語もある。

表 1: 感情別登録語句数

喜	怒	哀	怖	恥	好	厭	昂	安	驚
42	19	46	28	17	23	112	23	19	4

4.2 感情生起表現候補の収集

感情生起表現候補を抽出する際、4.1節で作成した種用感情表現辞書に登録した表現を含む文をコーパスから取り出す。取り出した文を南瓜(2)を用いて構文解析する。その結果から感情表現の係り元2文節を抽出する。また、その抽出した表現の末尾が「のが」、「ことが」であるものだけを感情生起表現候補とする。これは取り出した文を観察した結果、表現候補の末尾が「のが」、「ことが」である時、係り元の動作や事柄が感情の対象である場合が多いことが分かったためである。このため、本稿で定義した感情生起表現が多く取れることが予想できるのでこの処理を加える。また、「ことが」はそれ自身で1文節なので係り元を2文節抽出するときの文節としてカウントしない。つまり、「ことが」が表現末尾であった場合は感情表現の係り元3文節を感情生起表現候補とする。図1に抽出例を示す。

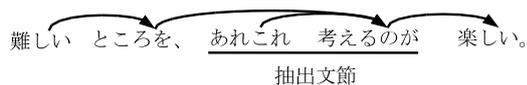


図1: 感情生起表現の抽出処理例

図1のような文の構造の場合、係り元2文節は「ところを、考えるのが」、「あれこれ考えるのが」の2パターン取れるが、本稿では感情生起表現を文中で連続した表現に限定したため、この場合「あれこれ考える」だけを生起表現候補として取り出す。また例5に「のが」を含む生起表現候補、例6に「ことが」を含む生起表現候補の例を示す。

例5) わいわい やる のが

例6) のびのびと 書ける ことが

4.3 感情生起表現候補のフィルタリング

フィルタリングは以下の手順で行う。また、4.2節で抽出した感情生起表現候補において2文節の前の文節を1文節目、後ろの文節を2文節目としている。

Step 1 1文節目の先頭の形態素が「連体詞」、「名詞-非自立」、「名詞-数」の場合は感情生起表現候補としない。これは感情生起表現は2文節の表現であると定義したためである。また本稿では「NがV(する|した)」といった基本的なものを対象としており、複雑な意味情報や係り受け情報は考慮しないことにする。そのため、2文節の表現を対象にする。また、感情生起表現候補である2文節より前の文節から情報を補完する処理を行っていないため、「こうした」等の連体詞を考慮しない。

Step 2 2文節目の先頭の形態素が「形容詞-自立」、「動詞、名詞-サ変接続」、「名詞-形容動詞語幹」の感情生起表現候補を抽出する。2文節目は感情表現の直接の係り元なので感情生起の原因が書かれている可能性が高い。そのため動作を表すような表現である場合はそうでない場合よりも感情生起表現であることが予想できる。

5 評価実験

5.1 コーパス

実験に使用したコーパスは我々が2004年に収集したWebコーパスと2005年に同じ手法で収集したWebコーパスの2種類を使用した。

- ・2004コーパス : 約400MB、約486万文
- ・2005コーパス : 約1GB、約898万文

5.2 実験方法

感情の分類ごとに得られた感情生起表現の候補群からランダムに50個取り出し判定した。感情生起表現候補が50個に満たない分類は抽出した全ての候補を判定する。被験者3人に抽出した表現とその表現を抽出する際に用いた種用感情表現辞書の感情分類を並べて提示し、評価を以下の4段階に設定した。

- 評価A: 「一般的に提示している感情を生起する」
- 評価B: 「人によっては提示している感情を生起する」
- 評価C: 「提示している感情以外の感情を生起する」
- 評価D: 「表現が感情を生起する要因とならない」

2人以上が選んだ評価をその表現に対する評価とし、3人が別の評価を選んだときは評価Dとした。例外として被験者の評価がA、B、Cに別れた時はその表現が感情を生起する表現であるということは分かるため評価Cとした。

5.3 感情生起表現抽出精度

表2にコーパス別の抽出精度、表3に感情ごとの抽出精度を示す。また、例7に(喜)、例8に(哀)、例9に(厭)の抽出例を示す。

例7) (喜)

- { 公共料金が安いことが }
- { 自然とのふれあいが }

例8) (哀)

- { 出番がすくないのが }
- { はかなく消えてしまうのが }

例9) (厭)

- { 展開が読めてしまうのが }

{ しっくり こなかったのが }

表 2: コーパス別抽出精度

	A[%]	B[%]	C[%]	D[%]
2004 コーパス	21	39	3	37
2005 コーパス	15	25	4	56
総合評価	16	28	4	52

表 3: 感情別抽出精度

	2004 コーパス		2005 コーパス		総合	
	A[%]	A+B[%]	A[%]	A+B[%]	A[%]	A+B[%]
喜	32	64	12	50	16	53
怒	13	63	14	47	14	50
哀	44	62	24	58	22	57
怖	16	54	14	44	20	48
恥	10	50	0	14	2	22
好	0	68	4	28	3	36
厭	28	62	26	52	26	54
昂	50	80	47	60	47	64
安	4	40	6	12	6	18
驚	6	69	2	22	3	32

6 考察

6.1 抽出表現数の推移

本稿で得られた感情生起表現候補の抽出表現数は 2004 コーパスでは 2131 個、2005 コーパスでは 8043 個であった。また、この 2 つの結果の異なり数は 10059 個となった。2004 コーパスの約 2.5 倍の 2005 コーパスから約 4 倍の生起表現候補が抽出できている。これは 2005 コーパスが 2004 コーパスに比べて感情表現を多い割合で含んでいるといえる。(喜)の感情表現を含む文は 2005 コーパスは 2004 コーパスの約 3.4 倍となっていた。

また、2 つの表現抽出数の合計と異なり数に大きな差が無い。このことから、コーパスサイズの増加に伴い、感情生起表現の抽出数も飽和せずに増加することが確認できた。田中ら [2] は結合価パターン中の約 5600 個に情緒属性を付与している。パターンからフレーズに拡張した場合、膨大な量の表現が予想される。本稿で定義した感情生起表現はパターンではなくフレーズであるため多くの感情生起表現が存在することが予想できる。

その他に、得られた感情生起表現には感情別にしたときに抽出数に大きな差が生じていた。もっとも多く抽出できた感情は(厭)で 3076 個となり、もっとも抽出数が少ない感情は(恥)で 27 個であった。コーパス

中で(厭)の表現を含む文は 11775 文に対し、(恥)の表現を含む文は 750 文であったことから、これはコーパス中のテキストに感情による偏りがあるためと分かる。

6.2 提案手法についての検討

本稿の提案手法の特徴は、コーパスから取り出す情報が係り受け関係と品詞情報と末尾の字面というシンプルな点である。しかし、種として感情表現を使っているため感情表現を含まない文は処理対象にならない。このため、Web 上の掲示板やブログ等、人の感情や評価が一般的なテキストよりも多く含まれる文章を利用することが望ましい。

また、表 2 の評価結果に示すように、評価 C の値が 4[%] となった。これより、提案手法で感情ごとに得られた感情生起表現は他の感情を生起する表現を抽出しづらいことが分かる。ここからさらに精度を上げるためには、4.2 節の感情生起表現候補を収集する部分で感情表現の係り元 2 文節の頻度情報によってフィルタリングすることが考えられる。本実験では感情生起表現候補の約 9 割が頻度 1 の感情生起表現であったのでこの処理を加えず実験を行った。

さらに本稿で対象としなかった原文で文節が連続しない表現は例 3 の「何かをもらう」と言う表現は「誰に」という情報が感情の生起に関わっている可能性は高い。今後検討の余地がある。

6.3 パタン辞書との違い

本研究は関連研究で使われているパタンによる取得手法や辞書構築とは違ったものを提案している。これは、パタンのように一般化を行わない手法でも、大量にある Web 上の文書を用いることで多くの感情生起表現が取れるのではないかと考えたためである。さらに、フレーズで取得することで「あれこれ考える」といった特徴的な表現が取れる反面、抽出の精度と再現率が下がってしまう。最近、Web 上では掲示板を用いて多くの人がいろいろなものに対して書評したり、Blog として日記形式で主観的な感情を綴っている。こういった主観的な文章から抽出することはより多くの感情生起表現が得られ、精度上昇の要因になると予想できる。

6.4 抽出誤りについての検討

得られた感情生起表現の中で評価 D と判定された表現のほとんどが感情生起表現の情報不足であった。例 10)

{ どんどん 出てくるのが }

原文：普段食べられないものが注文するとどんどん出てくるのがとても幸せでした。

例 10 は (喜) の感情表現を含む文から抽出され評価実験において評価 D と判定された表現である。原文から分かるようにこの文章では「普段食べられないものが出てくる」という事象が書き手の感情を生起している。このような表現は本稿の手法で取得することは難しい。しかし、「普段～ない」という表現は感情の生起と関係が強いと考えられるため、特殊な表現に対応する処理を加えることで取得は可能である。

その他の問題に、人によってうれしいや悲しいと感じる事象が異なる点がある。一般的にうれしいや悲しいと定義できるものは存在するが本稿では個人差が生じず同じ感情を生起する表現と生起する感情に個人差が生じるような表現に差を見つけることができなかった。

- ・個人差がなく同じ感情を生起する表現
プレゼントをもらう (喜), 葬式に参列する (哀)
- ・生起する感情に個人差が生じるような表現
プロ野球を見る (好 厭), 壇上で発表する (昂 恥)

6.5 今後の展望

感情を推定する際、本稿で構築を試みた感情生起表現辞書は主観で書かれた文から抽出した表現辞書なので Blog 等の日記や映画のレビュー等、個人の主観で書かれた文章に対応し易いことが予想される。今後は Blog やレビューの記事を感情表現辞書と本稿で抽出した感情生起表現辞書を用いてどの程度感情が推定できるのかという実験を行いたい。

7 感情表現との共起についての追加実験

感情表現と共起しやすい感情生起表現は 5.2 節の評価 A や評価 B になりやすいという予想に基づき追加実験を行った。抽出した感情生起表現を Google(3) を利用しフレーズ検索する。そして、検索結果のテキスト下の文章から感情生起表現を含む文を抽出する。検索件数を 50 件とし、抽出した文を南瓜を使って構文解析する。クエリとした感情生起表現の係り先に感情表現があるかを判定した。しかし、係り先に多く感情表現が現れた表現と被験者による評価と比べた結果、この実験で感情表現が係り先に多く出現した表現のほとんどが評価 A や評価 B になっていなかった。このことから感情表現と共起しやすいということが感情生起表現の抽出精度の向上には関係しないことが言える。以下に実験の成功例と失敗例を示す。

例 11) (喜) 育っていくのを見るのが

例 11 は成功といえる例である。この表現の場合、評価実験で A と判定され追加実験である Google の検索結果においても、検索した 50 件に対して係り先が得られた 42 件中 34 件が感情表現であり高い割合を占めている。

例 12) 予定がないのが (哀)

例 12 のような表現の場合、書き手の状態によっては別の感情が生起される。書き手が普段予定に空きがない人の場合、この例は (喜) を生起する可能性が高くなるはずである。しかし、そういった情報は 2 文節で補完することはほぼ不可能である。追加実験の結果は検索した 50 件に対して係り先が得られた 41 件中 24 件が感情表現で、このうち (哀) の感情表現が 23 件であったことから (哀) の感情生起表現と考えられる。しかし、評価実験では被験者の意見が A、B、C に別れたため評価 C となった。

8 結論

本稿では感情表現を種として感情生起表現を Web コーパスから抽出する手法を提案した。結果として、約 1.4GB のテキストから約 1 万個の感情生起表現を約 45% の精度で抽出することができた。

謝辞

本研究の一部は、平成 17 年度 長岡技術科学大学 学長裁量経費 研究助成 (A) 「小学校 2 年生の国語問題を解く」によって実施した。

使用した言語資源及びツール

- (1) 形態素解析器「茶筌」, Ver.2.3.3, 奈良先端科学技術大学院大学 松本研究室,
<http://chasen.naist.jp/hiki/ChaSen/>
- (2) 構文解析器「南瓜」, Ver.0.52, 奈良先端科学技術大学院大学 松本研究室,
<http://chasen.org/~taku/software/cabocha/>
- (3) サーチエンジン Google,
<http://www.google.co.jp/>

参考文献

- [1] 角谷由貴, 森田和宏, 泓田正雄, 青江順一: 電子メールの重要文抽出と分類, 言語処理学会第 11 回年次大会発表論文, pp.321-324, 2005.
- [2] 田中努, 徳久雅人, 村上仁一, 池原悟: 結合値パターンへの情緒生起情報の付与, 言語処理学会第 10 回年次大会発表論文, pp.345-348, 2004.
- [3] 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.
- [4] 中村明: 感情表現辞典, 東京堂出版, 1993.