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

2007－NL－180（6）
2007／7／24

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

山本 麻由† 土屋 誠司†† 黒岩 眞吾†† 任 福継††

† 徳島大学 大学院 先端技術科学教育部 〒770-8506 徳島県徳島市南常三島町2-1
†† 徳島大学 大学院 ソシオテクノサイエンス教育部
E-mail: †{yamamoto,tsuchiya,kuroiwa,ren}@is.tokushima-u.ac.jp

あらまし 感情に関わる研究において，言語データに発話者の感情を表すタグ（感情タグ）を付与した感情コーパスの構築が望まれている．しかし，人手で作成するには多くのコストを要する．そこで本稿では，感情コーパス作成の自動化を目指し，文中の語に基づいたナイーブベイズによる感情分類手法を提案する．Webから収集した学習データを用いた評価実験により提案手法の有効性を確認する．

キーワード ナイーブベイズ分類, 自動分類, 感情コーパス

Emotion Classification for Emotion Corpus Construction

Mayu YAMAMOTO†, Seiji TSUCHIYA††, Shingo KUROIWA††, and Fuji REN††

† Graduate School of Advanced Technology and Science, The University of Tokushima
2-1 Minamijosanjima,Tokushima 770-8506, Japan
†† Institute of Technology and Science ,The University of Tokushima
E-mail: †{yamamoto,tsuchiya,kuroiwa,ren}@is.tokushima-u.ac.jp

Abstract In this paper, we aim to develop Emotion corpus automatically using Naive Bayes Classifier. Emotion corpus is language data with emotion tags. Language data is the corpus which made by the sentences that we collected from web. Emotion tag stands for emotion of the people who wrote the sentences at the time. At first, we put emotion tags on the language data we collected. Next,we classify the language data using the Naive Bayes Classifier based on this data set, and I confirm the effectiveness of the method.

Key words Emotion Corpus, Naive Bayes Classifier, Automatic classification

1. まえがき

ユーザとスムーズな対話が行えるシステムには，ユーザの感情を理解したり，システムの感情を表現する感性情報処理が必要である．感性情報処理の分野では大規模な言語データを用いるアプローチが増えており，感情コーパスも必要とされるデータのひとつである[1]．感情コーパスとは，言語データに感情情報を持つタグ（感情タグ）を単語，文等に付与したコーパスである．感情コーパスを構築する主な目的は，人間の感覚に近いデータである感情コーパスを用いることでシステムが人間らしい感情表現を可能にすることにある．

感情コーパスを構築する際の問題として，感情は常に一意に決まるとは言えないということがある．感情の判断は，判断する人間の意見に依存することが考えられ，また周辺状況によっても意見は変化するといえるからである．この問題を解決するために，通常は数名での主観評価を行い，大多数の意見が一致した感情を正解とすることが多い．しかしながら，大規模なコーパスを構築することを考えると，複数名による主観評価を行うことは多くのコストを要する．そこで我々は感情コーパス構築のコストを削減するための手法について提案する．提案する手法はタグ付与の自動化手法である．合わせて，1人の主観評価で作成した感情コーパスの利用可能性についての調査も行う．

文に任意のタグを自動的に付与するという問題は，文を任意の感情に自動的に分類する問題と同じと考えられる．そこで我々は文の感情生起に大きく関わるものが文中の語すべてであると仮定し，分類手法として，文書分類に高い性能を発揮するナイーブベイズ分類[2][3]を使用する．一般的にナイーブベイズ分類は学習データを必要とする．今回は学習データ用のコーパスはwebの掲示板から収集した．日々増加し続けるwebは膨大な量を必要とするコーパスの資源として相応しいと考えたためである．また，本研究における感情コーパスとは，言語データに発話者の感情を表す感情タグを付与したコーパスとする．文1文に対して1つの感情タグの付与を行う．この収集し

たコーパスへ1名の主観評価によって感情タグの付与を行い学習データとし，ナイーブベイズ分類が感情の現れている文の分類に有効であるかを確かめるための実験を行う．ならびに1名の主観評価で作成した揺れのあるコーパスと数名の意見を元に作成されたやや平均的な意見のコーパスを比較し，それぞれのコーパスの特徴について考察する．

以下，第2節でナイーブベイズ分類について，第3節で作成した感情コーパスについて，第4節で評価実験について述べ，第5節でまとめる．

2. ナイーブベイズ分類

ナイーブベイズ分類について説明する．ナイーブベイズは，過去の事例をもとに未知の文書があらかじめ与えられているどのカテゴリに属するかを決定する分類手法である．

まず，分類単位であるコーパス中の1文を文書 d とみなす．文書 d はいくつかの単語 w で構成されている．

各感情カテゴリを $C_i = C(c_1, \cdots, c_i)$ とし，$d_j = d(w_1, w_2, \cdots, w_j)$ とする．すると d_j が分類されるべき感情カテゴリは，事後確率 $P(c_i|d_j)$ を最大化するようなカテゴリとなる．これは以下の式で求められる．

$$\begin{aligned}\cdot\ \hat{c} &= \underset{c_i}{\operatorname{argmax}} P(c_i|d_j) \\ &= \underset{c_i}{\operatorname{argmax}} P(c_i|w_1, w_2, \cdots, w_k) \\ &= \underset{c_i}{\operatorname{argmax}} P(w_1, w_2, \cdots, w_k|c_i)P(c_i)\end{aligned}$$

また，各カテゴリのもとで単語は独立に生起すると仮定し，

$$P(w_1, \cdots, w_k|c_i) = \prod_{k=1} P(w_k|c_i) \quad (1)$$

とする．

したがって，次式によって文書 d の感情カテゴリ c を決定することができる．

$$\hat{c} = \underset{c_i}{\operatorname{argmax}} P(c_i) \prod_{k=1} P(w_k|c_i) \quad (2)$$

$P(c_i) = \frac{\text{各カテゴリ } C_i \text{に含まれる文数}}{\text{コーパスの全ての文数}}$ である．$P(w_k|c_i)$ は次の様に求められる．

$$P(w_k|c_i) = \frac{N_i}{N} \quad (3)$$

$N = C_i$に含まれる総単語数，$N_i = C_i$において w_iが現れる回数である．これらの値は収集したコーパスから容易に算出することができる．

2.1 ゼロ頻度問題

単語 w が学習データ中に1度も現れない場合，w の出現確率 $P(w_k|c_i)$ は0と推定される．文の確率は複数の単語列の確率の積から算出するため，この場合，文中のどれか1つの単語の確率が0だと，文全体の確率も0となってしまう問題が発生する．この問題はゼロ頻度問題と呼ばれている．

この問題を避けるため，w の出現確率を求める際に単語の出現回数を補正した値を用いること（スムージング）が行われる．

表1 コーパスの一部

Table 1 a sample of the corpus

	文	感情
1	どーなるんでしょうか？	不安
2	ちょっと他人事ではないわー	不安
3	早く週末にならないかなぁ。	希望・期待
4	マッサージ行きたいなぁ。	希望・期待
5	宜しくお願い致します。	平静
6	みなさ〜ん、こんばんわ。	平静
7	本当に、うざかった !!!	嫌悪
8	…耐えられないっ	嫌悪
9	きもい、うざい、さいあく！	嫌悪
10	イヤぁ、ホント良かったです。	喜び
11	転職決まったんですね♪	喜び
12	なんかほのぼのしてて、好きだなぁ。	喜び
13	死ぬほどむかつく !!	怒り
14	あ〜腹立つことばっか !!	怒り
15	ったくふざけんじゃねーぞ !!	怒り

今回は単純な加算法であるラプラス法 [3] を用い，スムージングを行った．

加算法では，単語の出現確率を求める際に，出現回数 N_i に一律に一定の値を加える．出現回数の補正値として1を加える手法をラプラス法と呼ぶ．ラプラス法での出現確率は，V = コーパスの全ての文中に現れる異なる単語数 としたとき，次のように求められる．

$$P(w_k|c_i) = \frac{N_i + 1}{N + V} \quad (4)$$

3. コーパスの収集

実験に用いるデータは人手で収集した．yahoo 掲示板(注1)から書き手の感情が表れていると思われる発話文を収集し，1文ごとに感情タグ付けを行った．タグ付けは1文につき1名で行った．作成したコーパスの具体例を表1に示す．

使用する感情は6種類である．内訳は，Table 2 に示した5種類および感情のない状態を表す“平静”である．感情を選ぶ際には,Ekman の基本感情 [4] を参考にした．コーパスの感情カテゴリごとの文数，1文あたりの平均形態素数，カテゴリの形態素総数，カテゴリの形態素異なり数の内訳を Table 3 に示した．コーパスの総数は7212文であり，全部で7616のユニークな形態素が使用されていた．最も文数の多いカテゴリは，“怒り”であり，次いで“喜び”，“嫌悪”，“平静”と続く．統計的手法を用いるため文数の少ないカテゴリでは分類精度があまり良くない事が予想されるが，感情を考える上で必要なサンプルとなると考えた．1文あたりの平均形態素数はおよそ10語だった．“喜び”は平均8.20語と短めであり，”平静”は11.6語と感情間での文の長さにやや差が見られた．

(注1)：http://messages.yahoo.co.jp/

表 3　コーパスの規模

Table 3　Detail of the corpus

感情	文数 (割合)	平均形態素数	総形態素数	固有形態素数
全体	7212	10.0	72124	7616
不安	610(8.5%)	10.4	6343	1339
希望・期待	767(10.6%)	10.5	8077	1615
平静	1030(14.3%)	11.6	11985	2459
嫌悪	1161(16.1%)	9.7	11218	2186
喜び	1273(17.7%)	8.2	10378	2224
怒り	2371(32.9%)	10.2	24123	3300

表 2　感情の種類

Table 2　Emotions

喜び	怒り	嫌悪	希望・期待	不安

4. 評価実験

実験について述べる．実験前の準備として，前章で述べた感情コーパスをテスト用データと学習用データに分けることとする．テストデータは，各感情につき 20 文をランダムに抽出した計 120 文，学習データは残る 7092 文とした．

4.1　テストデータの作成

ナイーブベイズ分類に対する正解を決定するために，テスト用データとして抽出した 120 文について，1 文につき 5 名の回答者に対して評価アンケートを実施した．回答者の負担を軽くするため，120 文を 4 分割し，1 名につき評価を行う文は 30 文とした．よって，アンケートには 20 名が参加した．

評価の方法は，対象となる文を読んでもらい，その文を書いた者の感情と思われるものを 6 つの感情から 1 つ選ぶという形式で行った．そこで 5 名中 4 名以上の選んだ感情が一致した 99 文を最終的なテスト用データとし，選ばれた感情をナイーブベイズ分類に対する正解判定の基準とした．

4.2　実験結果

実験結果として正解率，精度 (Precision)，再現率 (Recall) を求めた．正解率は 76.8%であった．正解率とは，全体の事例数のうち正解 (5 名中 4 名以上の意見によりある感情カテゴリ C と判断された事例) とナイーブベイズ分類の出力が一致した数である．

$$\text{正解率} = \frac{\text{正解と判断された事例とナイーブベイズ分類の出力との一致数}}{\text{全事例数}}$$

今回の実験では，全 99 文中 76 文のナイーブベイズ分類の出力と正解が一致した．精度と再現率を表 4 に示す．精度とは，ナイーブベイズ分類によりカテゴリ C と分類された事例数のうちアンケートでもカテゴリ C が正解と判断された事例数の割合である．

$$\text{精度} = \frac{\text{正解と判断された事例とナイーブベイズ分類の出力との一致数}}{\text{ナイーブベイズ分類によりカテゴリ } C \text{ と分類された事例数}}$$

また再現率は，アンケートにより正解と判断された事例数のうちナイーブベイズ分類でもカテゴリ C と分類された事例数の割合である．

表 4　実験結果：精度および再現率

Table 4　Precision and Recall

感情	精度	再現率
不安	75.0%(6/8)	40.0%(6/15)
希望・期待	**87.5%(7/8)**	43.75%(7/16)
平静	75.0%(15/20)	83.33%(15/18)
嫌悪	76.5%(13/17)	92.86%(13/14)
喜び	71.4%(15/21)	93.75%(15/16)
怒り	80.0%(20/25)	**100.0%(20/20)**

$$\text{再現率} = \frac{\text{正解と判断された事例とナイーブベイズ分類の出力との一致数}}{\text{アンケートによりカテゴリ } C \text{ が正解と判断された文数}}$$

カテゴリごとの精度は，“希望・期待”がもっとも高く 87.5%であり，次に“怒り”の 80.0%，“嫌悪”の 76.5%と続く．

5. 考　　察

5.1　学習データについて

テストデータ 120 文のうち 99 文についての評価が，1 名の主観評価と 5 名中 4 名以上の主観評価とで一致した．単純に学習データもテストデータと同じ割合で評価が一致すると考えると，この学習データの精度は 82.5%といえる．また前章で述べた通り，この学習データを使用したナイーブベイズ分類の正解率は 75.8%(75/99) であった．線形に正解率が上がると仮定すると，学習データを精度 100%で作成することができれば，ナイーブベイズを使うことでおよそ 91.8%の割合で人間らしく感情を分類できるという予想ができる．

5.2　分類誤りについて

各感情カテゴリの再現率を見ると，“怒り”，“喜び”，“嫌悪”，“平静”の 4 つのカテゴリの再現率は 80%以上であるのに対して，“不安”“希望・期待”の再現率は 40%程度とかなり低い．再現率が低いということは，ナイーブベイズは“不安”“希望・期待”の 2 つを分類しにくいと言える．

ナイーブベイズは学習データ中の出現単語数などから分類カテゴリを推定するため，学習データが少ないと分類に支障を来す．実際にこの 2 つのカテゴリのコーパス数は他のカテゴリに比べて少ない．このことから，ナイーブベイズ分類では各カテ

ゴリとも1000文程度のデータ量は必要だと考えられる.

5.3 今後の課題

今後はさらなる精度の向上を目指すために，文中のすべての語を素性として使用するのではなく，ルールを作り感情分類に有効な語を選ぶ必要があるだろう．具体的には感情および品詞ごとに出現頻度の高い語を取り出し，感情生起の傾向を分析し，ルールを作成することを考えている．1語のみを使用するのではなくバイグラムやトライグラムを使用することや，文末や文頭など特定の位置に出現する語を考慮する必要がある．すでにある感情辞書に収録されている語を素性として利用することも考えられる．

次に学習データについては，1名で作成したコーパスを使用しても76.8%の正解率を出すことができた．よって，7〜8割程度の成功率が求められるシステムに使用するためのコーパスは1名で作成したもので十分であろう．前節で述べた通り，学習データの精度が上がると，分類の精度も上がると予想できる．そのため，8割以上の高い精度での分類が求められるシステムでは複数人による厳密なコーパスを使用するべきであろう．

最後に，再現率の低さとコーパスのデータ数の関係について考える．今回作成した感情コーパスは，収集した際の文数のまま特に何の手も加えていない．よって今後もコーパスの収集の際には感情ごとのデータ数の偏りは起こりうる．ということは，はじめからデータ数が少ないカテゴリがあるという前提で分類を行うべきだろう．つまりデータ数の少ないカテゴリでも再現率を上げる何らかの工夫が必要となる．

6. おわりに

本論文では感情に関する研究に必要なコーパスの作成を自動化する手法のひとつとして，文中の語に基づいたナイーブベイズによる感情分類手法の検証をおこなった．評価実験の結果，全体の正解率はおよそ76.8%であり，ナイーブベイズが感情の分類にある程度有効であることが示された．また，1名で作成したコーパスでも学習データとして有効であることがわかった．今後は用途によって複数名で作成したコーパスと使いわけることが考えられる．

文　献

[1] 徳久良子, 乾健太郎, 徳久雅人, 岡田直之, 規模とコストを考慮した感情タグつき言語コーパスの作成方法, 電子情報通信学会総合大会ワークショップ「心を持つロボット, 対話するロボット」講演論文集, 2001

[2] 阿部倫子、田中久美子、中川裕志, コメントを用いた映画の分類, 情報処理学会 NL 研究会 NL-150, pp.105-110, 2002

[3] 北 研二, 言語と計算 (4) 確率的言語モデル, 東京大学出版会, 1999.

[4] Ekman, P. ,An Argument for Basic Emotions, In N. L. Stein and K. Oatley (eds), Basic Emotions, 169-200, Hove, UK: Lawrence Erlbaum, 1992

付　録

参考として，分類の誤り例およびテストデータのうち回答者の評価が一致せず正解感情が決まらなかった文を掲載する．

表 A·1　ナイーブベイズ分類の誤り例
Table A·1　Examples of classification's error

	文	アンケート結果	ナイーブベイズ分類結果
1	車内びしょびしょになっていなければいいけど。	不安	平静
2	もう誰に何を頼っていいのかわかんない …	不安	怒り
3	うちの会社いったいどうなってしまうんだ …	不安	嫌悪
4	明日はお休みなので、天気だといいなぁ。	希望・期待	喜び
5	お二方とも大丈夫なのだろうか。。。	不安	平静
6	いい結果が出ますように ‥	希望・期待	不安
7	会社で健康診断はやってますか？	平静	怒り
8	今の内に帰ってお洗濯したいっ！	希望・期待	怒り
9	とっても、ブルーなんです	不安	嫌悪
10	人間関係で毎日へこんでます。	不安	喜び
11	私はドコモの携帯です。	平静	嫌悪
12	胃癌とかだったらどうしよう。	不安	希望・期待
13	きも〜〜〜〜〜〜〜〜〜〜い！	嫌悪	怒り
14	声が届くといいなあ。	希望・期待	不安
15	危なくてハラハラしちゃいます。	不安	平静
16	というか、なったらいいな〜♪	希望・期待	喜び
17	こーーーんなに簡単にできるんだ！♪	喜び	怒り
18	ご旅行の間、お天気に恵まれますように。	希望・期待	喜び
19	一人だけ休んで、の〜んびりと過ごしたいッス!!	希望・期待	嫌悪
20	母は大分回復してきましたが、まだ、何があるかはわからないです。	不安	平静
21	ローランドゴリラもマウンテンゴリラも生息地はアフリカです。	平静	喜び
22	やっぱツーリングしてぇ!!	希望・期待	怒り
23	あー　給与の桁が1個多かったらなぁ。	希望・期待	嫌悪

表 A·2　回答者の評価が分かれた文例

Table A·2　Examples that evaluation of respondents did not accord each other

	文	アンケート評価内訳 (5 名)						1 名の評価
		怒り	喜び	嫌悪	希望・期待	不安	平静	
1	気をつけてね。	0	0	0	2	2	1	不安
2	久々に会う面子だし、楽しみだわーーー！	0	3	0	2	0	0	喜び
3	道も狭かったような、、、	0	0	1	0	3	1	不安
4	あーーーーーーーーっ、ヤダ !!!!!!!!	3	0	2	0	0	0	嫌悪
5	もうすぐボーナスですねっ★☆	0	2	0	3	0	0	喜び
6	ああ、早くひとりぼっちになりたい…。	0	0	0	3	2	0	希望・期待
7	家族共々是非行きたいです。	0	0	0	3	0	2	希望・期待
8	掲示板でワイワイやればいいのに。	0	0	3	1	0	1	希望・期待
9	なにもないことを祈ります。	0	0	0	3	2	0	不安
10	良かった良かった。	0	3	0	0	0	2	喜び
11	はっきり言って、消えて欲しい。	2	0	3	0	0	0	嫌悪
12	あーストレスたまるー	1	0	3	0	0	1	嫌悪
13	あほらしくてやってられたものじゃない！	3	0	2	0	0	0	怒り
14	台風の直撃で大変ですね	0	0	1	0	3	1	不安
15	みなさんおはようございます。	0	1	0	1	0	3	平静
16	愛想も糞もなく、まことに感じが悪い！	2	0	3	0	0	0	嫌悪
17	あ〜〜〜っ　もう　明日から仕事じゃん	0	0	2	0	3	0	嫌悪
18	私はホテル内で働いていた時期がありました。	0	0	1	1	0	3	平静
19	落ち着かないぃ	0	0	2	0	3	0	不安
20	奥様　おだいじに	0	0	0	3	0	2	不安
21	ここはのどかで良いですよね〜。	0	2	0	1	0	2	喜び