品詞 n-gram による反響表現の抽出

NHK 放送技術研究所

小早川 健 (こばやかわ たけし)
kobayakawa.t-ko@nhk.or.jp

宮崎 勝 (みやざき まさる)
miyazaki.m-fk@nhk.or.jp

藤井 真人 (ふじい まひと)
fujii.m-ii@nhk.or.jp

八木 伸行 (やぎ のぶゆき)
yagi.n-iy@nhk.or.jp

1 はじめに

放送局に寄せられる反響を計算機によって分析する研究を行っている。寄せられる反響は、アンケートによる選択肢型の設問への回答から、電話や自由記述欄への回答まで多岐にわたるが、選択肢型は、集計などにより比較的容易に分析を行うことが出来るため、自由記述型の分析に焦点を当てている。

計算機による文書分類は、典型的には、文書に出現する単語の頻度を数えて特徴量を構成する方法 (Bag of words) により、その特徴量を分類ないしクラスタリングする [1]。一方、最近では、評判分析の研究も多く行われていて、評判を表す表現の抽出や分類の研究もある [2, 3, 4]。いずれの場合でも、分類/クラスタリングのタスクに応じた特徴量の選択は重要な要素のひとつであり、分類に寄与しない言語表現を特徴量から除外することで分類/クラスタリング性能の向上が期待される。

評判分析における文書分類は、キーワードとなる体言を分類する一般の文書分類と違い、評判を表現する文の述部の用言に着目して分類する。用言の主な用法は、文の述部と他の体言や用言の修飾であることから、品詞の系列に着目し、他の部分を修飾しているかを判定することで、文の述部として使われている用言を抽出することが可能であると予想できる。そこで、本研究では、人手による分類作業で重要となった用言表現を教師として学習した反響表現の抽出器を用いて、反響分析用の文書分類器の性能改善を試みる。

1.1 文書分類タスク

タスクは、視聴者に見立てた被験者 125 人に教養娯楽番組を見せて、その番組に対する感想を自由記述形式で書いてもらい、その記述を分類するものである。試聴した番組は表 1 にある通りの 4 番組で、主に番組 1 を試聴した時の感想を分類した結果を報告する。1 人の作業者がすべての番組について同一のクラスタでクラスタリングを行い、これを分類の正解とする。クラスタのラベルと番組 1 における構成を表 2 に示す。被験者 1 人あたり 4〜10 個の事柄についての記述があり、それぞれを別の意見として扱った。計算機で扱う場合は、クラスタ数 8 を既知とした文書分類器の学習を行い、その分類器が正しいクラスタに分類するかどうかで分類性能を測定する。

表 1 自由記述による感想の対象となる放送番組

	教養娯楽番組の題材
番組 1.	現代医師の海外ボランティア活動の紹介
番組 2.	海外音楽史上の人物の紹介
番組 3.	現代芸人の地域密着活動の紹介
番組 4.	日本史上の文化・習慣の解説

表 2 クラスタのラベルと番組 1 の構成

	クラスタのラベル	構成数 (構成比)
a.	肯定的な意見	384 (41.9%)
b.	否定的な意見	35 (3.82%)
c.	番組を見て考えたこと	260 (28.4%)
d.	番組を見て知ったこと	69 (7.53%)
e.	番組への要望	95 (10.4%)
f.	番組への質問	15 (1.64%)
g.	その他の意見	23 (2.51%)
h.	意見でないもの	35 (3.82%)
	計	916 (100%)

反響の例として、番組 1 を試聴した時の感想の中からいくつか拾い出して表 3 に示す。クラスタは、人手による分類によって付与された表 2 のクラスタを表す。

表 3　反響例

	文章	クラスタ
ア.	緑の大地を少しずつでも取り戻す様子に感動。	a.
イ.	本当に人間はここまで人の為に出来るのか!!!	a.
ウ.	正直あまり興味深く見ることができなかった	b.
エ.	海外での活動は大変だと思った	c.
オ.	日本人でもこのように平和的に活動している人が多いんだなと思った。	d.
カ.	現地で働く日本人青年たちの、仕事以外の日常生活も見てみたかった。	e.
キ.	なぜアフガニスタンなのか	f.
ク.	日本ではかなえられない夢があるの言葉にも複雑な思いがしました	g.
ケ.	水があると、牛と子供が先に集まってくると話をしていた。	h.

2　文書分類手順の概要と特徴量

文書分類手順の概要を図 1 に示す。茶筌 [5] を用いて文書の形態素解析を行い、文書を文書特徴量に変換する。反響表現の抽出アルゴリズムを通じて、文書分類器に入力される。

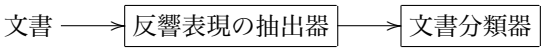

図 1　文書分類手順の概要

2.1　反響表現の抽出のための特徴量

反響分析の抽出に用いる特徴量は、形態素解析の結果で得られる品詞の系列を用いる。例えば、本タスクの番組 1 に現れる単語は 2,183 種類であり、品詞は 58 種類と比較的に少ない。そこで、品詞 n-gram の $n \leq 3$ 程度までならば、学習データが不足すること (データスパースネスの問題) なく反響表現の抽出器を学習することが出来るため、実際に本研究でも品詞 3-gram までを用いる。

2.2　文書分類器のための特徴量

文書分類器に用いる特徴量は、形態素解析の結果で得られる単語を用いる。これは、単語が出現した回数で表現する Bag of words モデルである。品詞の系列を用いると分類の弁別性能が十分に得られないことが予想され、単語の系列を用いるとデータスパースネスの問題が生じることが予想されるため、特徴量に用いることは避けた。

3　反響表現の抽出器

3.1　ナイーブ・ベイズ分類器とスムージング

(離散) 特徴量を A, 分類クラスを C とすると、特徴量 A に対して最も可能性の高いクラス $\hat{C}(A)$ は、ベイズの定理を用いて、

$$\hat{C}(A) = \operatorname*{argmax}_C P(A|C)P(C) \tag{1}$$

と表され、ベイズ分類器と呼ばれる。特徴量が K 次元の場合でも、同様のベイズ分類器

$$\hat{C}(A_1, \ldots, A_K) = \operatorname*{argmax}_C P(A_1, \ldots, A_K|C)P(C) \tag{2}$$

が可能である。これに対し、各特徴量の独立性を仮定すると、ナイーブ・ベイズ分類器

$$\hat{C}(A_1, \ldots, A_K) = P(A_1|C) \cdots P(A_K|C) \cdot P(C) \tag{3}$$

が得られる。

i 番目の学習データが特徴量 a を持ち、クラス c に分類される時、変数 $d_{a,c}^{(i)}$ は 1 を取り、その他の場合は 0 を取るとする。また、$\mathcal{A}$ は特徴量 A の取り得る値を要素とする集合である。確率分布は学習データから最尤推定すると、

$$P(A|C) = \frac{\sum_i d_{a,c}^{(i)} \delta_{A,a} \delta_{C,c}}{\sum_{\alpha \in \mathcal{A}} \sum_i d_{a,c}^{(i)} \delta_{\alpha,a} \delta_{C,c}} \tag{4}$$

となる[*1] が、実際には、調整すべきパラメター λ を用いた以下のスムージング[*2]を適用する。

$$P(A|C) = \frac{\sum_i d_{a,c}^{(i)} \delta_{A,a} \delta_{C,c} + \lambda}{\sum_{\alpha \in \mathcal{A}} \left(\sum_i d_{a,c}^{(i)} \delta_{\alpha,a} \delta_{C,c} + \lambda \right)} \tag{5}$$

反響表現の抽出器では、このナイーブ・ベイズ分類器を用い、分類クラス C の取り得る値は、抽出する/しないの 2 値としている。また、表 4 のように、特徴量 A として品詞 n-gram($n \leq 3$) を用いる。品詞 1-gram は着目した品詞単独の出現、前方-品

[*1] (4) 式の中で、δ は次式で定義される。

$$\delta_{i,j} = \begin{cases} 1 & (i = j \text{ のとき}) \\ 0 & (i \neq j \text{ のとき}) \end{cases}$$

[*2] これは、Laplace のスムージングを拡張したもので m-推定と呼ばれる。

詞 2-gram は着目した品詞とその前方に出現する品詞の組み合わせ、後方-品詞 2-gram は着目した品詞とその後方に出現する品詞の組み合わせ、品詞 3-gram は着目した品詞とその前後に出現する品詞の組み合わせを表す。図 2 の例では、着目した品詞を実線で囲んであり、その場合の特徴量を表 5 に示す。

表 4　特徴量

特徴量	種類
A_1	品詞 1-gram
A_2	前方-品詞 2-gram
A_3	後方-品詞 2-gram
A_4	品詞 3-gram

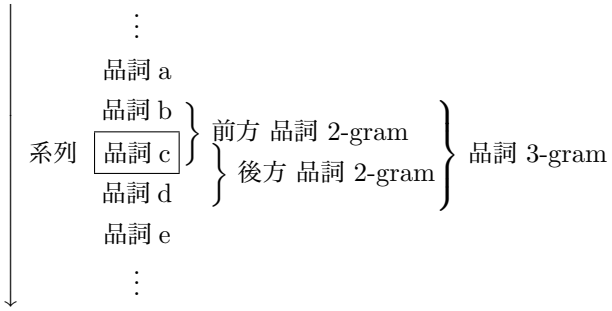

図 2　品詞系列の例

表 5　例の場合の特徴量

特徴量	図 2 の場合の品詞 c の特徴量
A_1	品詞 c
A_2	品詞 b+ 品詞 c
A_3	品詞 c+ 品詞 d
A_4	品詞 b+ 品詞 c+ 品詞 d

3.2　反響表現の抽出実験

前小節に述べた反響表現の抽出器を用いて、番組 1 の反響表現の抽出実験を行う。表 6 は、スムージング係数 λ を 0.5 に設定した時の、特徴量の組み合わせと抽出性能の関係を示す。closed 評価は番組 1 への反響をすべて評価データとし、そのすべてのデータを学習に用いた場合の評価である。open 評価は、番組 1 のデータの 1/10 を評価、9/10 を学習に用いた 10-fold 交差検定の場合の評価である。open domain 評価は、番組 1 への反響をすべて評価データとし、番組 2〜4 への反響を学習データに用いた場合の評価である。

番組 1 では、全データに対する抽出すべき部分の割合が 20.6(%) であることから、open domain 評価の場合に、本手法はほとんど効果がないことがわかる。一方、同一ドメイン内では、特徴量の組み合わせとして品詞 1-gram と後方-品詞 2-gram(A_1+A_3) が最も高性能であることがわかる。

図 3 は、学習データを評価に用いた場合のスムージング係数 λ と反響抽出率の関係を表す。使用した特徴量は品詞 1-gram と後方-品詞 2-gram(A_1+A_3) である。スムージングを行わない $\lambda = 0$ の場合と比較して、$0 \leq \lambda \leq 1$ の範囲で誤抽出が減少していることがわかる。

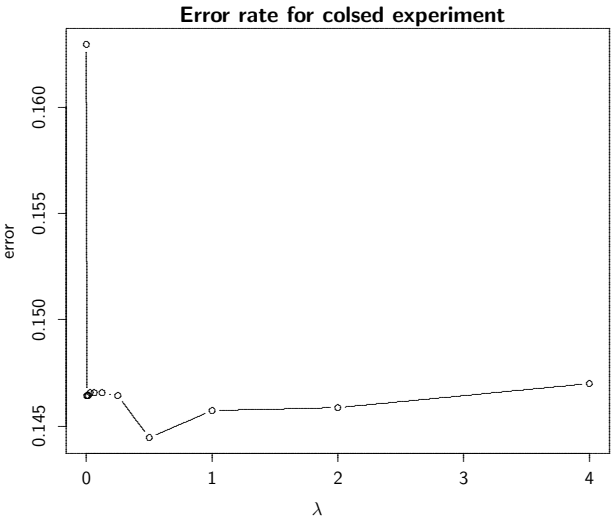

図 3　スムージング係数 λ と分類性能の関係 (closed 評価)

4　文書分類実験

反響表現の抽出器を前段に用いて、文書分類実験を行う。反響表現の抽出器は入力文書に対するフィルターのように動作し、抽出すると分類された単語のみを次段の文書分類器に渡す。反響表現の抽出器ですべてを抽出しないと判定された文書は、フィルターとして動作しないので、すべてを抽出して文書分類するという例外処理を行っている。分類器には多項式カーネルによるサポートベクターマシンを用いる。

表 7 に文書分類実験の誤分類率を示す。closed 評価は、番組 1 の反響表現の抽出実験において closed 評価で得られる出力を学習データに用い、同データをすべて評価データに用いた場合の評価である。open 評価は、番組 1 の反響表現の抽出実験におい

表 6　反響表現の抽出器の誤抽出率

特徴量	closed 評価	open 評価	open domain 評価
$A_1 + A_2$	16.9(%)	17.7(%)	19.9(%)
$A_1 + A_3$	13.4(%)	14.5(%)	20.2(%)
$A_1 + A_2 + A_3$	16.3(%)	17.6(%)	20.4(%)
$A_1 + A_2 + A_3 + A_4$	14.3(%)	16.8(%)	21.3(%)

表 7　文書分類実験の誤分類率

実験の種類	closed 評価	open 評価	open domain 評価
反響表現抽出器なし	0.0(%)	39.5(%)	57.1(%)
手作業による反響表現抽出	2.0(%)	31.9(%)	60.7(%)
反響表現抽出器 ($A_1 + A_2$) あり	7.4(%)	41.9(%)	58.2(%)
反響表現抽出器 ($A_1 + A_3$) あり	4.6(%)	38.4(%)	64.8(%)
反響表現抽出器 ($A_1 + A_2 + A_3$) あり	1.4(%)	39.3(%)	69.3(%)
反響表現抽出器 ($A_1 + A_2 + A_3 + A_4$) あり	1.5(%)	38.1(%)	56.4(%)

て open 評価で得られる出力を用いた 10-fold の交差検定での評価である。open domain 評価は、番組 2〜4 の反響表現の抽出実験において open 評価で得られる出力を学習データに用い、番組 1 の反響表現抽出実験において open 評価で得られる出力を評価データに用いた場合の評価である。

実験結果は反響表現の抽出実験と同様の傾向を示す。open domain 評価では、表 2 での最大構成比が 41.9(%) であることから、有効な文書分類器の誤分類率は 58.1(%) 以下であり、本手法はほとんど効果がないことがわかる。反響表現抽出器が用いる特徴量によって性能のバラツキがみられるが、有意な差ではなく誤差の範囲と思われる。一方、同一ドメイン内の open 評価では、反響表現抽出器がない場合の 39.5(%) に比べ、反響表現抽出器がある場合が誤分類が減少している場合が見られる。しかし、その改善効果は、期待していたほどではなく、僅かなものである。手作業による反響表現抽出は、有意に文書分類性能を改善することから、反響表現が上手に抽出できれば有効であることは確かめられた。本研究では品詞 n-gram のみを用いたが、文末からの位置情報など、より多くの情報を用いた反響表現抽出器の可能性は残されている。

5　おわりに

品詞 n-gram を特徴量としたナイーブ・ベイズ分類器による反響表現の抽出器を実現し、その抽出性能を実験した。同一ドメイン内の実験では、反響表現の抽出器はごく僅かの効果があり、その抽出器を用いると文書分類の分類性能が向上する場合が見られるという実験結果が得られた。一方、放送に対する反響に限定した場合でも、異なるドメインの反響を抽出する実験はほとんど効果がなかった。これは、品詞 n-gram だけでは、情報が極めて限定的であることを示唆している。

また、単純な Bag of words モデルでは文書分類器の性能がかなり低いことから、今後は、単語よりも大きな単位で反響表現を解析できる技術に取り組んでいきたい。

参考文献

[1] M. W. Berry Ed.: “Survey of Text Mining”, Springer (2004).

[2] 小林, 乾, 松本, 立石, 福島：“意見抽出のための評価表現の収集”, 自然言語処理, **12**, 3, pp. 203–222 (2005).

[3] 大塚, 内山, 井佐原：“自由回答アンケートにおける要求意図判断基準”, 自然言語処理, **11**, 2, pp. 21–66 (2004).

[4] 乾, 村田, 内山, 井佐原：“表層表現に着目した自由回答アンケートの意図に基づく自動分類”, 自然言語処理, **10**, 2, pp. 19–42 (2003).

[5] 浅原, 松本：“統計的日本語形態素解析に対する拡張 HMM モデル”, 情報処理学会研究報告 2000-NL-137, pp. 39–46 (2000).