E-058

放送に対する反響抽出の課題

Issues on Extracting Opinion for Broadcasting

小早川 健†
Takeshi S. Kobayakawa

1 はじめに

放送局に寄せられる番組への反響を分析する研究を行っている。先行研究 [1] では、反響を典型的なカテゴリ (表 1) に分類する問題と捉え、機械学習的な手法を適用している。

表 1 分類カテゴリ

a.	肯定的な意見	e.	番組への要望
b.	否定的な意見	f.	番組への質問
c.	番組を見て考えたこと	g.	その他の意見
d.	番組を見て知ったこと	h.	意見でないもの

反響の種類で分類するタスクとは別に、反響の対象によって分類するタスクも考えられる。一般的な評判分析では、対象は商品名のように有限かつ固定的であるため、対象によって分類するタスクは容易である。しかし、放送に対する反響は、多種多様な表現で対象を表現し、また、放送番組毎に対象が異なるという問題を含んでいる。本報告では、放送に対する反響の対象を放送言及部分と呼ぶことにし、一般的な評判分析に対して、放送番組に対する評判を分析する場合の課題を明らかにする。

2 放送言及部分を特定する必要性

2.1 一般的な評判分析

一般的な評判分析は、文から<対象, 属性, 属性値, 評価>という 4 つ組で意見を抽出するが、属性値と評価の区別が困難な場合があるため、属性値と評価をあわせて評価値とし、<対象, 属性, 評価値>という 3 つ組の抽出問題として捉える [2]。

一般的な評判分析での<対象>と<属性>は、全体と部分の関係にあると考えるとわかりやすい。(正確には披修飾と修飾の関係にある。) <属性>が無い場合は全体に対する評価であるのに対し、<属性>がある場合は属性で全体を制限して部分を表している。属性による制限は 0 回、または、1 回に限定されてはなく、以下の架空の例を考えると、多重にすることが出来ることがわかる。下線部は、対象と属性に相当する部分である。

例. 車がいい
例. 車の色がいい
例. 車の内装の色がいい
例. 車の後部座席の背もたれの色がいい

† NHK 放送技術研究所

2.2 放送番組に対する評判分析

放送に対する反響について調べるために、ある特定の放送* に対する反響から、放送言及部分を人手で抽出した。そこには、<対象>と<属性>の枠を越えた、以下の実例のような複雑な表現が見受けられる。下線部は、放送言及部分である。

例. 「相手の立場に立って考える」という中村医師のことばが心に残った。
例. 人間は愛するに足るという言葉に励まされた。
例. 不毛の大地を緑に変える水の大切さを知らされた。

このように、<対象>-<属性>の枠を越えた複合的な表現になっているものは、反響文全体数の約 1/3 もの多数を占める。

複合的な表現になっている理由のひとつには、反響表現が、具体的な物以外の抽象的な事柄を指していることがあげられる。参考までに、日本語語彙体系 [3] を基にした独自のシソーラス (図 1) を定義し、放送言及部分にどのような単語が出現しているのかを集計 (図 2) したら、「事」の下に分類される表現が事例の半数近くに及んでいることがわかった。「事」の下に分類される表現は、複合的な放送言及部分を多く構成している。

以上の事から、放送に対する反響文は、対象そのものを特定するのに名詞の抽出だけでは不十分であり、難しいタスクとなっていることがわかる。そこで、放送言及部分を機械的に特定することを試みる。

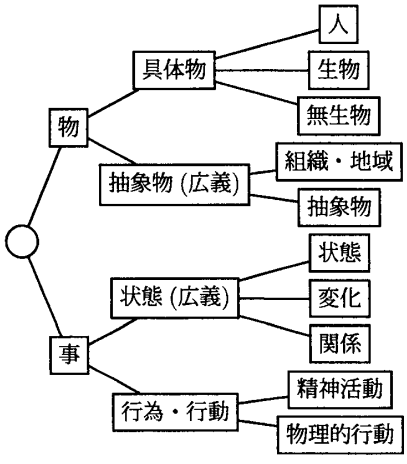

図 1 放送言及部分を分類するためのシソーラス

* 2006 年 7 月 24 日放送 NHK 教育テレビ「知るを楽しむ」という番組。中村医師という人物の活動に焦点を当てた番組で、戦争で荒廃したアフガニスタンで井戸堀りをして水の供給源を開拓しながら診療所を開設するボランティア活動に励んでいる。中村医師は番組中のインタビューで、「相手の立場に立って考える」、「人間は愛するに足る」、「真心は信ずるに足る」という発言をしている。

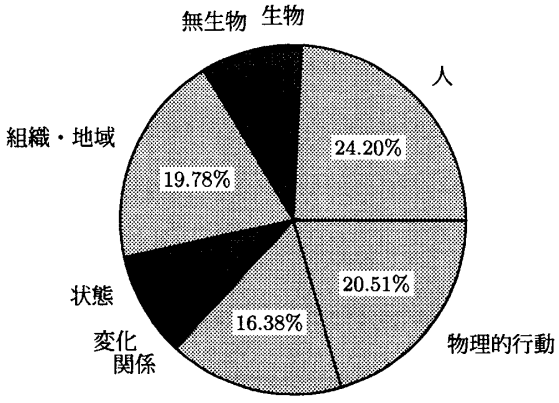

図 2　放送言及部分の内訳

3　放送言及部分を特定する手法

3.1　提案法

t 検定による方法 [4] を拡張して一連の表現を検出し、他の放送回から検出された表現との差分をとることによって、放送言及部分を特定する方法を提案する。

まず、[4] の t 統計量を n-gram に拡張する。

$$t = \frac{\bar{\chi} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{p_n(w^1 \cdots w^n) - p_1(w^1) \cdots p_1(w^n)}{\sqrt{\frac{p_n(w^1 \cdots w^n)}{N}}} \quad (1)$$

次に、これを用いた下記のアルゴリズムにより、10-gram までの表現を検出する。

アルゴリズム

- for $n = 10$ to 2
  - n-gram の中から 1 つ w_n を選ぶ
    - w_n の出現確率 p_n を計算する
    - w_n を 1-gram に分解し、その出現確率の積 p_1 を計算する
    - 式 (1) により、p_1 と p_n から t を計算する
    - t が有意水準を乞えていたら w_n を候補として採択する
    - $N > n$ なる N に対して w_n が w_N の部分文字列ならば w_n を候補から除外する
    - w_n が候補ならば、出力する
  - w_n のループを続ける
- n のループを続ける

そして、その放送回に特徴的な表現だけを抽出するために、他の放送回で検出された表現との集合差分をとる。最終的に差分で残った表現を元の反響文と照合することによって、放送言及部分を特定する。

3.2　実験

提案法を用いて、有意水準を 0.5(%) と設定して実験を行った。1 放送番組に対する 916 文の反響文から検出された (集合差分をとる前の) 表現の一部を表 2 に示す。灰色の背景で塗りつぶされた表現が放送言及部分である。他の放送回で検出された表現との差分をとった後 (表 3) では、放送言及部分が多く残っていることがわかる。

全反響文のうち、放送言及部分を正しく特定できた数 (A) と、特定洩れを起こした数 (B) と、誤特定を起こした数 (C) を用いて、

$$\text{precision} = \frac{A}{A+C}, \quad \text{recall} = \frac{A}{A+B} \quad (2)$$

で定義される精度を、形態素と文字を単位として計測した結果を表 4 に示す。

表 2　検出された表現

出現回数	t	反響表現
7	2.65	「相手の立場に立って
10	3.16	が印象的だった。
16	4.00	だと思った。
12	3.46	人と人との
11	3.32	考えさせられた。
11	3.32	に感動した。
10	3.16	のではないか
8	2.83	と思いました。
7	2.65	は愛するに足る

表 3　集合の差分で残った表現

「相手の立場に立って
人と人との
は愛するに足る
水の大切さ
日本の若者に
の大切さを
NGO の

表 4　精度

単位	形態素	文字
precision	50.94 (%)	52.71 (%)
recall	18.57 (%)	17.98 (%)

4　おわりに

放送に対する反響を分析する場合に、反響の対象である放送言及部分を特定することが難しいタスクになっていることを指摘した。検定を用いた方法により反響言及部分を特定する方法を提案し、実験によりその精度を確認した。実験によると、precision よりも recall がかなり低くなっており、今後の改良が期待される。

参考文献

[1] 小早川, 宮崎, 藤井, 八木：“品詞 n-gram による反響表現の抽出”, 言語処理学会第 13 回年次大会 (2007).

[2] 小林, 飯田, 乾, 松本：“照応解析手法を利用した属性-評価値対および意見性情報の抽出”, 言語処理学会第 11 回年次大会 (2005).

[3] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林：“日本語語彙体系 CD-ROM 版”, 岩波書店 (1999).

[4] C. D. Manning and H. Schütze: “Foundations of Statistical natural Language Processing”, The MIT Press (1999).