TV番組に対する自由回答文の印象抽出システム
—インターネットアンケート調査による自由回答文の解析—

月出 奈都子 石崎 俊
{hitachi,ishizaki}@sfc.keio.ac.jp
慶應義塾大学 政策・メディア研究科

1 はじめに

インターネットを活用したテレビの視聴質調査であるResearchQ[1] [1]によって収集された自由回答文を分析し、大量の自由回答文を俯瞰するためシステム"Impression Extraction System(IES)"を開発した。まず、番組にとって良い意味合い、悪い意味合いの言葉を「印象語」とし、分類し辞書を作成した。IESは、その辞書と回答文の形態素解析結果を用い、文中の印象語を抽出し、また印象語の出現率を出力する。本稿で用いる回答文は、番組への5段階評価と同時に収集されているが、その評価値とIESによる印象語の出現率との間の相関について考察する。

2 データの収集方法

ResearchQのサイトでは、毎日、過去4日間(3日前、2日前、昨日、今日)の番組表を(19時以降)更新し、以下のような質問を用意している。

> 番組を見ようと思って選びましたか?放送された番組については、あなたが「見た」番組をチェックして質問にお答えください。
>
> 1. この番組を見ようと思って選びましたか?
> 2. この番組を見て、よかったですか?
> 3. この番組を集中して見ましたか?
> 4. この番組について意見・感想・質問があれば、どうぞ

3項目までの質問は、「そう思う」「まあそう思う」「どちらでもない」「あまりそう思わない」「そう思わない」の5段階評価となっている。また、それぞれ「集中度」「満足度」「期待度」と名付けられている。4つ目の項目は自由回答形式によって回答する。

[1]全国朝日放送株式会社(略称 テレビ朝日)マーケティング部と慶應義塾大学 環境情報学部 熊坂研究室が、1997年4月より始めた共同プロジェクト

3 IESシステム構成

IESでは、印象語辞書が用意されており抽出処理過程で参照される。

3.1 印象語辞書について

どの語を抽出するか決定するための「印象語辞書」を作成した。

TV番組にとって、文脈に左右されずにいつでも「良い意味合い」「悪い意味合い」である言葉を、それぞれ「プラス表現」「マイナス表現」とし、辞書に登録した。また、文脈によっては良い意味にも悪い意味にもなるものは、「ニュートラル表現」とする。例えば、「恐い」はホラー趣旨の番組には誉め言葉に成り得るが、番組によっては悪い意味の時もあるからである。

データを集約させるために、類似した意味の言葉をグループ分けした。例えば、「良い」「よい」「好い」「善い」「いい」は、表記の違いはあっても意味の違いが小さい。さらに回答文中の「すばらしい」は回答者が「よい」と感じているのとは大差ないため、同じグループとする。また「良い」の反意語は「悪い」で、否定の意が付加されると「良くない」であり、「悪い」と意味的な近さをもつ。この際に、「良くない」という表現と「悪い」がグルーピングされるように、「良い」と「悪い」は同じグループに登録してある。

3.2 印象語辞書の作成方法

「計算機用日本語基本形容詞IPAL (Basic Adjectives)[3] (以下IPAL)を元に印象語辞書を編集した。まず、IPALの形容詞、形容動詞の見出し語136語すべてを取り上げた。次に、表記、異音同語、派生形容動詞、派生形容詞、同義語、類

義語、反義語の項目に記載されている語すべてを印象語辞書の見出し語として登録した。

ある見出し語に対して、他の項目に記載されている語はすべて同一のグループとみなした。IPALでは見出し語「つらい」の類義語には「くるしい」が、逆に見出し語「くるしい」の類義語には「つらい」が登録されている。このような場合は、見出し語の「つらい」「くるしい」に含まれる、その他の項目の語が、印象語辞書では一つのグループとなる。このように、関係する語を繋ぎ合わせることで、より大きなグループへと編集し、結果として198グループとなった。

また、「評価」「快不快」の項目を利用し、印象語辞書におけるプラス、マイナス、ニュートラル表現を決定した。

IPALを元に編集した印象語辞書を、さらに収集された4555件の自由回答文を参考にして整備した。回答文に出現した動詞と形容詞のうち、「印象語」となるものを登録した。例えば「見る」「思う」といった語の出現頻度は非常に高いのだが、回答者の印象としては関係が薄いため、登録はしない。一方、「見たい」「共感(する)」「嫌い」といった語は、重要な語とみなして登録した。自由回答文を参考にピックアップされた語は、「分類語彙表[4]」「角川類語新辞典[5]」などの、資料、辞書をグループ分けの参考にした。

3.3 抽出課程

まず、形態素解析システム「茶筌2.0b6」[2]により文中の形態素情報を得る。だが、抽出と集約処理の都合上、例えば以下のような場合に形態素情報を修整しておく。

- 「人間\関係」など、名詞の類が連続して出現している場合→「人間関係」
- 「共感\が\できる」、「感心\が\ある」といった「ある特定の語」+「助詞」+「動詞」→「共感(する)」「感心(する)」
- 「気\に\なる」、「気\に\いる」などの統合

修整された形態素情報を用いて、回答文中の動詞、形容詞の類ごとに項目に区切る。例えば、「この時間帯はあまり注目されていませんが、今回は出演者も豪華ですし、いいと思います」ならば、次のようになる。

否定形： 注目・ 時間帯
肯定形： 豪華： 今回, 出演者
肯定形： いい：
肯定形 思う：

そのように得られたものを印象語辞書に照合し抽出すると同時に、印象語のグルーピングする。自由回答文100件に対しての出現率、印象語群、名詞類の関連語を出力し情報の集約をする。表1にその一部を示す。

表1: 抽出後の例
ドラマ「恋の奇跡」への自由回答文376件の解析結果
「おもしろい」をはじめとする1グループの結果
出現率：回答文の数を100とした時の印象語の出現回数

出現率	印象語	関連語
24.5	おもしろい 面白い 興味 引き込む はらはら 楽しい たのしい 楽しむ たのしむ	11件, 展開; 6件, 今後; 4件, メイク; 3件, 反撃; 葉月; ストーリー; 変身. ぶり; 役; 2件, 葉月. 里緒菜; 話; 菅野; (以下省略)

また、プラス、マイナス、ニュートラル表現ごとに、IESへ入力された回答文の数を100とした時の印象語の出現率を計算する。

IESの処理過程では、形態素解析の修整をした後に、要望文の判定を行う。方法は、表2のような表層的な語や表現をもつ文に関してIESの処理過程から外し、要望文としてストックしておく。一つの文のなかに、複数の特徴が出現している場合も多い。

表2: 要望文の表層的な特徴

表現	例
ほしい(欲しい)	Ｊｒをたくさんテレビに出してほしい
たい	本当はもっと共感したいんですけど
ください(下さい)	違うテーマも検討して下さい
もうすこし(もう少し)	俳優をもう少し若い人にしてほしかった
もうちょっと	もうちょっと、どの年代のひとが見ても面白いって思えるお話にして欲しいです。
もっと	三上がもっと切れた役であってほしかった
活用語が仮定形の時	若い人たちが見ればおもしろいのかも知れませんが。

4 印象語出現率と5段階評価値との関係

第2節で紹介したように、ResearchQでは自由回答文は「集中度」「満足度」「期待度」の5段階評価と共に収集されている。そこで、番組ごとにIESによる印象語出現率、5段階評価値の平均[2]を求め、相関について考察する。

各番組をバラエティ、ドラマ、スポーツ、音楽、アニメ、ニュースの6つのジャンルに分類し、ジャンルごとで印象語出現率と5段階評価値の相関係数を求めた。その結果、例えばスポーツ番組における プラス表現と満足度の相関係数は0.559、散布図は図1である。

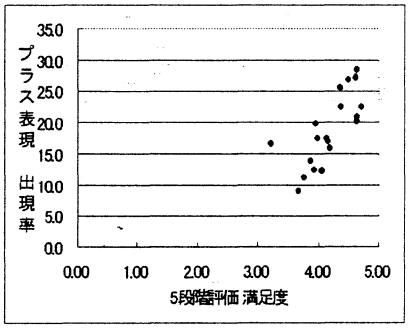

図1: プラス表現出現率と満足度：スポーツ番組
相関係数 :0.559

スポーツ番組のマイナス表現と満足度＼集中度の他、音楽番組におけるプラス表現＼マイナス表現と期待度にも、ある程度強い相関の傾向がある。一方で、その他のジャンルにおいてはそれほど顕著な相関は見られない（図2参照）。

原因として、スポーツ番組においては自由回答文の表現が他のジャンルに比べて、複雑ではないため、5段階評価とよく似た傾向になると考えられる。また音楽番組では、回答者の好きなアーティスト（出演者）に関する回答文が多く、そのため期待度との相関が強くなるのではないかと思われる。

他のジャンル、特にドラマにおいては、自由回答文から抽出される印象語の表現も多岐にわたり、また出現率の高低も番組ごとで大きく変化する。

[2]回答者全体の4分の3が5段階評価のみの回答で、自由回答文の件数は全体の4分の1程度である。ここでは自由回答文への記述のある回答者の5段階評価値のみを扱う。

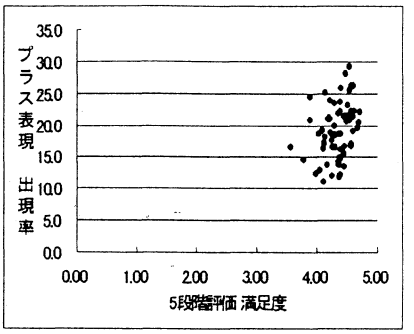

図2: プラス表現出現率と満足度：ドラマ番組
相関係数 :0.261

このため、自由回答文には「期待度」「満足度」「集中度」の他の評価軸となり得る可能性がある。

5 おわりに

TV番組にとって良い意味合い、悪い意味合いの言葉である「印象語」を中心に自由回答文の情報抽出を行うシステムIESを開発した。自由回答文が単純であるスポーツ番組においては、印象語の出現率と5段階評価値との対応が見られた。他のジャンルに関しては、番組ごとの分析を今後行っていく予定である。また現在、IES抽出結果と人手の要約の比較を行っている。

謝辞

データの利用を快諾していただきましたテレビ朝日マーケティング部ならびに慶應義塾大学熊坂研究室の皆様に心よりお礼申し上げます。

参考文献

[1] URL: http://fatimah.sfc.keio.ac.jp/RQ/

[2] URL: http://cl.aist-nara.ac.jp/lab/nlt/chasen.html

[3] 情報処理振興事業協会 技術センター, 計算機用日本語基本形容詞辞書 IPAL(Basic Adjectives),1990.

[4] 国立国語研究所, 分類語彙表, 大日本図書株式会社,1964.

[5] 大野晋 浜西正人:角川類語新辞典, 角川書店,1981.

[6] 那州川哲哉:テキストマイニング-膨大な文書データの自動分析による知識発見-, 情報処理 40 巻 2 号,1999.

[7] 乾裕子 内元清貴 井佐原均:文末表現に着目した自由回答アンケートの分類, 言語処理学会第4回年次大会発表論文集,1998.