OS-01c-1

係り受け情報を加味したオノマトペ感性抽出の試み

An Attempt to Extract Sentiment Value of Onomatopoeic Expressions using Grammatical Dependency

橋本 喜代太[*1] 岡田 真[*1] 竹内 和広[*2]
Kiyota Hashimoto Makoto Okada Kazuhiro Takeuchi

[*1] 大阪府立大学 Osaka Prefecture University
[*2] 大阪電気通信大学 Osaka Electro Communication University

Onomatopoeic expressions, particularly those in Japanese and Korean, are used to describe subjective impression on things and events, and thus they are indispensable for sentiment analysis. However, their extraction and estimation of their sentiment value is not trivial. In this study, we employed a shallow parsing, or superficial dependency, to adapt machine learning to different usages of an onomatopoeic expression with different sentiment values.

1. はじめに

オノマトペ(擬音語・擬声語・擬態語・擬容語・擬情語などを総称して本稿ではオノマトペと呼称する)は現実世界の音を模倣したり，行為・状態の様態を音表象したりする表現であり、日本語や韓国語など一部の言語においてはきわめて使用頻度が高いと同時に、その記述内容は話者・書き手の主観に負う面がきわめて大きい。つまり、これらの言語における感性情報抽出などでは欠くことのできない言語表現であると言える。

オノマトペはきわめて生産的かつ異形態も多く、日本語においてはひらがな、カタカナ、さらにはその入り混じりで表現されるために事前処理となる形態素解析においても十全に取り出せるとは限らず、その意味はもちろん感性情報についても事前に十分な辞書登録をするのは困難である。このため、従来からデータからのオノマトペの自動獲得や概念辞書の自動生成[奥村 2003]が試みられてきた。

その一方で、主にポジティブ、ネガティブという粗い感性評価の獲得を考えた場合、オノマトペには大別して 2 つの問題点がある。一つはオノマトペそのものの感性評価は主観性が強いために必ずしもそのポジティブ、ネガティブがはっきりしないケースが少なくないことであり、いま一つは行為・状態の様態を音表象的に記述する表現であるがためにそれが何に対する記述であるのかによってポジティブ、ネガティブがまったく異なってくるということである。前者については人それぞれの好みというものがある限り、まったく同じ行為・状態についてもその評価は異なってくるのが当然であるから、全面的な解決はできない。一方、後者については感性辞書の作成において大きな問題となる。また前者についてもある程度は後者が改善されることによって改善される可能性もある。

本研究においてはこの後者の問題を今後解決していくための予備的実験として係り受け情報を用いることによって改善を試みるものである。

2. 感性表現としてのオノマトペ

2.1 オノマトペの表す概念と感性値

オノマトペは主観性が強い一方、その主観性の多くが同一言語話者に共有されているという興味深い側面を持つ。例えば「ぼろぼろ」、「ぽろぽろ」、「ぽろんと」、「ぽろんと」といった類似表現の違いは外国人にとっては把握困難なことが多いが、日本語話者であればほぼ共通で微細な違いを把握できている。

しかしながら、その「共有された主観性」と「その価値判断をポジティブと捉えるかネガティブと捉えるか」はまったく別の問題であり、そこでは主観性が強いからこそ個人差も生じる。つまり、オノマトペ表現の概念/意味を獲得することと感性値を獲得することはイコールではない。本研究では後者を目的とする。

感性値の獲得を目的とする場合のもっとも基本的な手法はあらかじめ辞書登録単語に感性値を割り振っておき、それらとの共起頻度によって新たな語の感性値を獲得するというものである。また、文章の感性値はそうした感性値を持つ語の頻度の総和として計算されることになる。今回は前者に焦点を絞る。

2.2 バグオブワーズと係り受け

ある語の感性値を獲得する上で一般的なのはある長さの文字列中（文といった不定長を単位とすることもある）にあらかじめ感性辞書登録された語とどれくらい強く共起するかを計算するものである。この時、形態素（単語）の集合として文字列を定義するのがバグオブワーズであり、学習のためのデータ数を稼げる割に採用する手法等次第でそれなりの精度が出ることから頻用されるが、「麺はつるつるなんだけど、野菜のシャッキリ感がいまいちで、ちょっと微妙な感じ」といった例で「つるつる」、「シャッキリ」の感性値は「いまいち」、「ちょっと微妙」のネガティブさに引きずられて学習される危険性があることになる。なお、ここでも見られる逆接や譲歩、否定による感性値の反転も別の問題として存在する。

一方、係り受け情報を利用する場合、基本的には利用する共起語の数 i だけ対象語を分けて共起頻度を見るため十分なデータ数を揃えるのが難しい。また、どのような係り受け情報を利用するかによって結果は大きく異なってくることになる。本研究ではこの点についてデータの観察を通して、有効な係り受け情報を推定し、実験で検証する。

3. 係り受け情報の検討

本研究ではデータとしてオノマトペが頻用されやすい料理記述に焦点を絞り、レストラン等の口コミサイト、料理ブログ等から約百万字を収集したデータを利用した。

このデータについて約十万字について人手で観察し、オノマトペに分類できる語を確定し、そのうち 20 回以上の頻度がその

連絡先：橋本喜代太，大阪府立大学，599-8531 堺市中区学園町 1-1，hash@kis.osakafu-u.ac.jp

約十万字中に見られたもののみを今回のターゲットとすることとし(63 語)、形態素解析器に単語登録されていないものは予め品詞情報とともに登録をした(12 語)。

3.1 記述対象語との係り受け

オノマトペのほとんどはいわゆる形容動詞、副詞、サ変動詞のいずれかの品詞を取る。このため、記述対象語との係り受けは次の 3 種 5 類である。

- X が Y だ/Y する(主述関係 1)
- X が Y している (主述関係 2)
- Y(に) X する (連用修飾関係)
- Y な X (連体修飾関係 1)
- Y する X(連体修飾関係 2)

この点はもともと明らかなことなので特に議論しない。なお、「すっきり感」、「すっきり度」のようにオノマトペを評価尺度語化する用法については感性評価値を決めること自体が無意味なので除外している。

3.2 他文節との接続関係

今回の事前観察では、オノマトペは他の形容詞・形容動詞に比べても記述対象語との文中距離が隣接またはきわめて近いものが圧倒的多数を占めていた。これはデータが口語的なものであることも関係している。

一方、感性評価値推定のために必要な他の感性評価値付与語との関係できわめて特徴的だったのは次の 3 パターンである。

- 「おいしくってもうにっこにこ」

この例のようにオノマトペで文ないし、読点で次の独立した文につながる(「最後はスープをごくごく、おいしうございました」)といった例がよく見られる。これも口語的であることが反映されている。

- 「後味すっきりで、もう最高！」

この例はオノマトペ「すっきり」が使われている長文節は「で」で後続に順接しているが、明確に理由を述べるものとなっている。このようにオノマトペが使われている部分が理由を表すとき、「～なので」等の理由表現ではなく「～で/して(いて)」の順接表現が好まれる傾向があった。

- 「麺はしこしこでスープはこってり」

この例は並列だが、逆接となる場合も含めて、オノマトペによる記述では 2 つないし 3 つの連関するものが並列的・対照的に使われるものがよく見られた。もちろん単独で使用される例も多いが、こうした並列的利用は特にデータ数が多くないとき、感性評価値の推定に効果的に利用できると考えられる。

これら以外にもいくつかの特徴が観察されたが、今回は主にこの 3 点に着目した。

4. 実験と考察

前節での観察を踏まえて、次のような実験を行なった。まず各オノマトペをその記述対象語ごとに別用法とみなして w_1, w_2 のようにする。ただし、そのままではデータがスパースになりすぎるので、肉、野菜など同じカテゴリーに属するものは同じとみなした。ただし、麺についてはうどん、ラーメン、スパゲティと蕎麦、素麺の 2 タイプでは大きく評価が異なったため、この 2 タイプに分割した。

そのうえで、人手による観察の際に正解となる感性評価値を付与した十万字分を教師データとし、判断基準となる感性評価値付与済み表現としては「よい」、「いまいち」などの少数に限ったものを別途用意した。

この教師データで各用法に分類したオノマトペのそれぞれについて感性評価値がポジティブとなるかネガティブとなるかを SVM を用いて学習させ、それを残りの九十万字のデータ(各十万字の 9 セット)で検証した。

この結果、対照実験として行なったバグオブワーズとして係り受けを一切用いない場合に対して、13～24%の精度向上が見られた。一方、今回はオノマトペがどんな記述対象語に対して用いられているかに分けたわけだが、その結果、ほとんどのオノマトペについてはどの記述対象語でもポジティブ、ネガティブは同じであった。つまり、料理など特定のジャンルの場合、一つのオノマトペが記述対象語ごとに違った感性評価値を取るケースが稀であったということを示している。ただし、これについては料理というジャンルの特徴かもしれず、今後の検討が必要である。

5. 結語

本研究ではオノマトペの感性評価値を自動獲得するに当たって、その記述対象ごとにオノマトペの感性評価値が異なるという事実に対応するため、係り受け情報を活用してより精度の高い推定を行なうことを試みた。前節のようにそれは一定の成果を得たが、将来課題も多い。

まず、対象としたオノマトペの多くがその記述対象が何であるかに関係なく一意の感性評価値を持つことが確認された。一方、今回ターゲットとした料理関係ではそうでも他の分野ではそうではない、という事例もある。このため、オノマトペとその記述対象について改めて人手による丁寧な観察を試みたい。ただし、この点にも関わらず、本研究で特に着目した文節の順接・逆接を捉えるために係り受け情報を利用するという手法は有効である。

今回は推定のタネとなる感性評価値付与済み表現を最小限度に限るという試みで実験を行なった。これは特に料理の場合、一般の名詞表現等はポジティブ、ネガティブどちらの文脈でもよく登場するなど推定のタネとなりにくい一方、「よい」、「おいしい」など感性評価値がはっきりした表現が共起するケースが比較的多かったためである。この点についてもそれが料理というジャンルを超えて一般化が図れるかはのちの検討としたい。

以上、本研究では単純なレベルであるが、オノマトペの感性評価値を最小限のタネ語をもとに推定するにあたって、特徴的な係り受けを利用することを試みた。

参考文献

[奥村 2003] 奥村敦史, 齋藤豪, 奥村学: “Web 上のテキストコーパスを利用したオノマトペ概念辞書の自動構築, ” 情報処理学会 自然言語処理研究会 2003-NL-154-10, pp.63-70, 2003