

オノマトペの音象徴を利用した評判分析

五十嵐 沢馬[†] 笹野 遼平[‡] 高村 大也[‡] 奥村 学[‡]

[†] 東京工業大学 総合理工学研究科

[‡] 東京工業大学 精密工学研究所

igarashi@lr.pi.titech.ac.jp {sasano, takamura, oku}@pi.titech.ac.jp

1 序論

言語音が持つ印象のことを音象徴 (Sound Symbolism) と呼ぶ [1]. 我々はこの音象徴を日常で無意識的に活用しており, 音象徴を分析することは言語処理において有用であると考えられる. 例えば, 日本語のオノマトペは音象徴が顕著に表れる言語表現であり, オノマトペの評価極性を推定する場合に音象徴は有効であると考えられる. 評価表現辞書の構築に関する研究は数多くあるが [5], このような手法は, 新語, 未知語への対策が不可欠であり, オノマトペのように新しい語が生まれやすい表現には適していない. 日本語話者は, 未知のオノマトペに対して, その語感を頼りに解釈している. これは, オノマトペの形態的, 音韻的特徴が規則的であるためである. 特に音韻的特徴は音象徴と密接な関係にあり, この特徴を利用して既知のオノマトペと未知のオノマトペの紐付けを行うことで, 未知のオノマトペに対しても高精度に極性を推定することができると考えられる. 本研究では, オノマトペの音象徴を利用することで, 評判分析の精度が向上することを示すことにより, オノマトペの極性推定における音象徴の有用性を明らかにする.

2 関連研究

近年, 自然言語処理をはじめとする計算機科学の分野ではオノマトペの研究が盛んである. 提案手法と関連するものとしては, オノマトペの分類に関するものが挙げられる. その手法を大別すると, 文脈の情報に基づくものと, オノマトペの音韻的特徴に基づくもの, あるいはその両方の情報を用いたものである. 文脈の情報に基づくものとしては, 黒澤ら [4] の自己組織化マップを用いた手法や, Komiya ら [3] の階層型クラスタリングを用いた手法がある. 音韻的特徴に基づくものとしては, 加納ら [7] の研究があり, 文脈, 音韻両方の情報を用いた研究は Ichioka ら [2] の研究がある.

これらの研究の主な目的は, 人間のオノマトペ理解の支援である. 例えば, 日本語学習者にとってオノマトペは感覚的で理解が困難であるため, そのような人

のために, オノマトペの理解を支援することを目的としていることが多い. 一方, 本研究は, 一般的な評判分析のシステムを対象とし, オノマトペの音象徴を利用することで, より高精度の極性推定をすることを主な目的とする.

3 提案手法

提案手法は, オノマトペを含む評価文が与えられた時に, その文脈上でのオノマトペの評価極性を教師あり学習によって学習し, 推定するものである. 例えば, 入力の評価文として「彼はニコニコと笑っていた」という文が与えられた場合は「ポジティブ」, 「彼はゲヒゲヒと笑っていた」という文が与えられた場合は「ネガティブ」と出力する. 極性の推定には, Support Vector Machine (SVM) を用いる. 以下では, 分類に用いる素性について記述する.

3.1 音象徴に関する素性

オノマトペの音象徴は, そのオノマトペの発音から感じられると考えられる. オノマトペの発音を表す方法としては, いくつか考えられる.

1 つ目は仮名文字である. 仮名は表音文字であるため, 仮名自体も発音表現となる. しかし, 発音そのものを表すためのものではないため, 発音を表す素性として利用するにはやや不十分であると考えられる.

2 つ目は音素記号である. 仮名を子音と母音に分けて表現する方法であり, 「かたかた」は /katakata/, 「ことこと」は /kotokoto/ となる¹. これにより, 文字表記は異なっているが, 五十音の同じ行, 同じ段の音が考慮できるようになる. オノマトペの音象徴を表現する際, 一般的に用いられている表現である.

3 つ目は音声記号である. 音声記号は音素記号よりも厳密に発音を表現した体系である. 例えば, 音素では, 「は」の子音と「ふ」の子音のように, 異音の関係にある音も同じ音 /h/ と見なす. 一方, 国際音声記号 (IPA) では, 「は」の子音は無声声門摩擦音 [h], 「ふ」の

¹一般的に // で括った記号は音素記号を示す.

表 1: 素性として加える IPA の分類カテゴリ

1. 破裂音	7. 唇音	13. 歯茎音
2. 摩擦音	8. 舌頂音	14. 後部歯茎音
3. 鼻音	9. 舌背音	15. 硬口蓋音
4. 弾き音	10. 咽喉音	16. 軟口蓋音
5. 摩擦音	11. 両唇音	17. 口蓋垂音
6. 接近音	12. 歯音	18. 声門音

子音は無声両唇摩擦音 [ɸ] であり、別の音と見なす²。

この 3 つの表現のうち、音象徴をモデル化するために最適な表現はどれか、という点に関しては、4.1 節の実験で示す。なお、日本語の各仮名と IPA の対応付けは諸説あるが、提案手法では「日本語音声学入門 [6]」に基づいている。特に、日本語のイ段は、硬口蓋化によって拗音と同じ子音になることを明記しておく³。ガ行鼻濁など前後の音や、語頭・語末などの位置によって、音が変わるものは考慮していない。

提案手法では、音象徴素性を基本的に N-gram の bag-of-words として表現する。また、素性値は、各素性の出現数を uni-gram, bi-gram ごとに正規化し、出現確率を表す値とする。

N-gram 素性は、どのような音が使われているかを表す素性である。そのため、音同士の類似性は考慮されないが、IPA は各音に分類があり、音同士の類似性を考慮することができる。例えば、「ま」の子音 [m] (両唇鼻音) と「ば」の子音 [p] (無声両唇破裂音) は記号が異なるため N-gram 素性では独立に扱われるが、共に両唇音であるため、その類似性を考慮できる。したがって、提案手法では、IPA の子音の分類に基づき、表 1 に示す素性を 2 値で加える。表中の 1 から 6 は調音の方法 (調音法) の分類、11 から 18 は調音の場所 (調音点) の分類、7 から 10 は調音の場所の上位分類となっている。オノマトペに出現する子音の数は不定であるため、対象のオノマトペに一番最初に出現する子音 (第一子音) について、この素性を加えることとする。これは、第一子音が音象徴の主導的役割を果たしている、という Hamano[1] の言に基づくものである。

3.2 文脈に関する素性

オノマトペの極性は、音象徴だけでは決まらない場合がある。例えば、「頭ががんがんする」という文の「がんがん」はネガティブな極性を持っていると考えられるが、「仕事をがんがん進める」という文の「がんがん」はポジティブな極性を持っていると考えられる。

²一般的に [ɸ] で括った記号は音声記号を示す。

³例えば、「き」の子音は「きゃ」の子音 [kɨ] と同一になる。

つまり、同じオノマトペでも、その極性は文脈によって変わる場合がある。そこで、提案手法では、音象徴素性に加えて、文脈情報も考慮する。具体的には、入力されたオノマトペを含む文中の名詞、動詞、形容詞の bag-of-words を素性として加える。素性値は、文中での単語の出現確率とする。

4 実験

本研究では、2 つの実験を行う。1 つ目の実験は、文脈を考慮しない、オノマトペ単体での音象徴を利用した極性推定の実験であり、2 つ目の実験は、文脈を考慮した実験である。

4.1 音象徴を用いたオノマトペの極性推定

この実験は、音象徴の素性のみを用いて、文脈のないオノマトペ単体の極性推定をする実験である。また、3.1 節で述べた点についての確認を行うことも目的としている。すなわち、仮名文字、音素記号、音声記号のうち、音象徴の表現に最適なものはどれか、という点の確認を行う。

実験データは、「現代擬音語擬態語用法辞典 [8]」から作成した。この辞典には、1064 語のオノマトペに関して、表している意味と用法、用法ごとのイメージについて「プラス」「ややプラス」「プラスマイナスのイメージはない」「ややマイナス」「マイナス」の 5 段階で記述されている。このうち、用法によって極性が反転しない、プラス/ややプラスイメージを持つオノマトペ 225 種類とマイナス/ややマイナスイメージを持つオノマトペ 620 種類を選択し、前者をポジティブクラス、後者をネガティブクラスとした。極性が反転しないとは、プラス/ややプラスの用法と、マイナス/ややマイナスの用法のどちらか一方のみを持つものを指す。

4.1.1 実験設定と評価尺度

分類器は線形 SVM による二値分類器とする⁴。実験データを正例、負例それぞれをランダムに 10 分割し、そのうち 9 つを訓練データ、残り 1 つをテストデータとした。評価尺度は 10 分割交差検定による分類正解率 (Accuracy) とする。この実験では、3.1 節で述べた音象徴表現の方法が異なる表 2 に示す 5 つの素性設定を比較する。uni-gram, bi-gram, tri-gram は、それぞれの記号での N-gram 素性を示し、category は、3.1 節で述べた IPA の分類に関する素性を示す。I3C2 にお

⁴SVM の実装には SVM^{light} を利用した。

表 2: オノマトペ単体の極性推定で比較するモデル

モデル	使用する素性
K3	仮名 (uni-gram, bi-gram, tri-gram)
P3	音素 (uni-gram, bi-gram, tri-gram)
I3	音声 (uni-gram, bi-gram, tri-gram)
I3C	音声 (uni-gram, bi-gram, tri-gram), category
I3C2	音声 (uni-gram, bi-gram, tri-gram), category, 子音 bi-gram

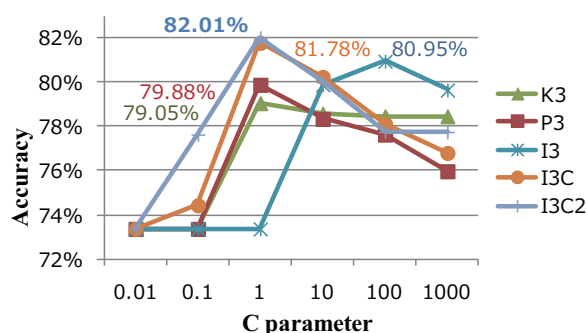


図 1: 仮名文字, 音素記号, 音声記号の音象徴性の比較

る子音 bi-gram は, オノマトペの音声記号列から子音のみを抜き出して作られる, 子音記号のみの bi-gram 素性である. これは, 子音同士の繋がりも音象徴に影響すると考えたために加えた.

4.1.2 実験結果

実験結果を図 1 に示す. 図の横軸は, SVM のパラメータ C の値であり, 縦軸は分類正解率である. C の値によって変動するが, 仮名文字より音素記号が正解率は高く, 音素記号より音声記号が正解率が高いことがわかる. 最も性能の良いモデルは I3C2 素性を用いたものであり, 文脈を考慮した実験では, このモデルの出力を素性として利用している.

4.2 オノマトペの音象徴を用いた文の極性推定

この実験では, オノマトペの音象徴を利用した文の極性推定をする. 文脈を考慮した分類を行うため, 文脈付きのデータとして「食べログ⁵」の口コミ文書を取得した. オノマトペを含む文の抽出には, 日本語形態素解析器 JUMAN Ver.7.0⁶を利用した. JUMAN にはオノマトペの自動認識機能があり, 辞書に載っていないオノマトペを解析することが可能である.

自動で抽出された文から, オノマトペ表現でない文と, 文脈とは関係のないオノマトペ表現の文, 否定語

表 3: 付与する正解ラベルの種類

ラベル	数	意味
ラベル 1	276	評価文がポジティブであり, 極性がオノマトペに依存 e.g. 「川がさらさらと流れる」
ラベル 2	76	評価文がポジティブであり, 極性がオノマトペに非依存 e.g. 「商品がどんどん売れる」
ラベル 3	170	評価文がネガティブであり, 極性がオノマトペに依存 e.g. 「川がごうごうと流れる」
ラベル 4	29	評価文がネガティブであり, 極性がオノマトペに非依存 e.g. 「株価がどんどん下落する」
ラベル 5	272	ニュートラル, すなわちポジティブ, ネガティブどちらでもない
ラベル 6	177	評価文から極性を判断できない

表現を含む文⁷などを取り除き, 残ったものから 50 種類のオノマトペについて, 各 20 文ずつ, 合計 1000 文を無作為に選択する. 正解ラベルの付与は, 2 名で独立に行った後, 両者の合議によって 1 つの事例につき 1 つのラベルに決定する. 付与するラベルを表 3 に示す. このような 6 つのラベルを付けている理由は, ラベル 1, 3 のように極性がオノマトペに依存している場合は, 文脈の素性は手掛かりにならず, 音象徴を考慮した提案手法の有効性がより顕著に表れると考えられるためである. このラベルのうち, ラベル 1, 2 をポジティブクラスのラベル, ラベル 3, 4 をネガティブクラスのラベル, ラベル 5 をニュートラルクラスのラベルとする. ラベル 6 は人間でも極性の判断ができない事例であるため, 実験データから除外する.

4.2.1 実験設定と評価尺度

未知のオノマトペに対する極性推定の評価を行うため, テストデータに含まれるオノマトペは, 訓練データに含まれない, という制約を付け, 実験データを訓練データとテストデータに分割する. さらに, ポジティブ, ネガティブ, ニュートラルのデータの割合が同程度になるように 10 個に分割し, 8 つを訓練データ, 1 つを開発データ, 残り 1 つをテストデータとする. 評価尺度は 10 分割交差検定による分類正解率とする. テストデータの分類時における, SVM のパラメータ C は {0.1, 1, 10, 100} の中から, 開発データを用いたテストで, 最も正解率が高くなる値を適用する. 分類するクラスは, ポジティブ, ネガティブ, ニュートラル

⁵<http://tabelog.com/>

⁶<http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN>

⁷否定語の自動判別の精度が本実験に与える影響をなくすため.

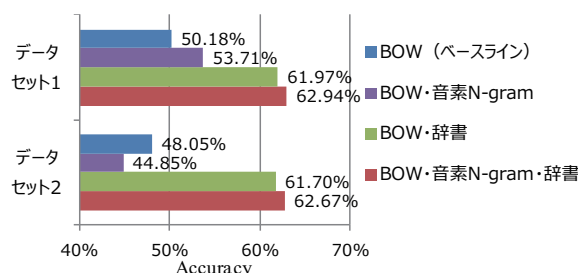


図 2: オノマトペの音象徴を用いた極性推定結果

の3値であり、one-versus-restによる3値分類をする。実験は、以下の2つのデータセットで行う。

データセット1 ラベル1から5の全データを用いたセット
データセット2 ラベル1, 3, 5のデータを用いたセット

データセット1は全データを用いたセットであるのに対し、データセット2は極性がオノマトペに依存し、文脈には依存していない事例に限定したデータセットであり、音象徴の素性の効果が有効であることを示すためのデータセットである。

ベースライン手法は、3.2節で述べた名詞、動詞、形容詞のbag-of-wordsモデル(BOW)とする。これは、音象徴を考慮しない従来手法となる。それに対し、提案手法は以下の素性を用いたものである。

音素 N-gram 音素記号の uni-gram, bi-gram, tri-gram による素性 (3.1 節)。

辞書 4.1 節の I3C2 素性で訓練された分類器の出力。

2 回目の辞書情報の素性は次のようにして得られる。予め 4.1 節で使用したオノマトペから、食べログの実験データに含まれるオノマトペを除いた 820 種類のオノマトペを用い、4.1 節と同様に 2 値分類器を訓練。素性は、評価対象のオノマトペを前述の 2 値分類器で分類し、出力の値が正なら 1、負なら 0 とする。すなわち、1 次元のバイナリ素性となる。

4.2.2 実験結果

データセット1の結果については、ベースラインの BOW モデルが 50.18% の正解率であったのに対し、BOW に加え、音素 N-gram、辞書情報の全素性を用いたモデルが最も高く、62.94% の正解率であった。マクネマー検定においても、1% 水準で有意差がある結果である。他の 2 つのモデルもベースラインより高い結果を示した。特に辞書素性は効果が高く、訓練データより規模の大きなオノマトペ辞書を用いて音象徴を学習したことが功を奏したと考えられる。

データセット2の結果については、ベースラインの BOW モデルが 48.05% の正解率であったのに対

し、データセット1と同様に全素性を用いたモデルが 62.67% の正解率であった。BOW に音素 N-gram を加えたモデルはベースラインよりも正解率が低くなった。これは、訓練データ中のネガティブクラスの事例が少なかったために、ネガティブな音象徴の学習ができなかったためと考えられる。辞書素性が有効に働いているため、音象徴素性自体は分類に寄与していると言える。データセット2の結果はデータセット1の結果と比較して、ベースラインと全素性を用いた提案手法の性能差が大きくなっており、極性がオノマトペに依存するようなデータでは、音象徴素性がより重要であることが確認できた。

5 結論と今後の課題

本研究では、オノマトペの音象徴を利用した文の評価極性を推定する手法を提案した。これまでの評判分析では、オノマトペの形態的、音韻的特徴に注目していなかったが、オノマトペの音象徴を用いることで、動詞、名詞、形容詞の bag-of-words のみを用いた従来手法より、音象徴を考慮した素性を加えた提案手法の方が高い分類正解率を得られることが分かった。今後の課題としては、より詳細な音象徴のモデル化や、オノマトペの共起情報などに注目した教師なしデータの活用が挙げられる。

参考文献

- [1] Shoko Hamano. *The Sound-Symbolic System of Japanese*. Tokyo: Kuroshio, 1998.
- [2] Kenichi Ichioka and Fumiyo Fukumoto. Graph-Based Clustering for Semantic Classification of Onomatopoeic Words. *Coling 2008: Proceedings of the 3rd Textgraphs workshop on Graph-based Algorithms for Natural Language Processing*, pp. 33–40, 2008.
- [3] Kanako Komiya and Yoshiyuki Kotani. Classification of Japanese onomatopoeias using hierarchical clustering depending on contexts. *Proceedings of the 2011 English International Joint Conference on Computer Science and Software Engineering*, pp. 108–113, 2011.
- [4] 黒澤義明, 竹澤寿幸. 自己組織化マップ SOM を用いた擬情語の分類比較. 人工知能学会 第 25 回全国大会, pp. 1C2-OS4b-7, 2011.
- [5] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, pp. 201–241, 2006.
- [6] 斎藤純男. 日本語音声学入門 改訂版. 三省堂, 1997.
- [7] 加納政芳, 戸本裕太郎, 中村剛士, 小松孝徳. 音響的特徴に基づくオノマトペの分類. 人工知能学会 第 25 回全国大会, pp. 1C2-OS4b-3, 2011.
- [8] 飛田良文, 浅田秀子. 現代擬音語擬態語用法辞典. 東京堂出版, 2002.