

語義情報の利用による感情・感覚に関連したオノマトペのクラスタリング

黒澤 義明

竹澤 寿幸

広島市立大学大学院 情報科学研究科

{kurosawa, takezawa}@ls.info.hiroshima-cu.ac.jp

1. はじめに ～ オノマトペ

オノマトペは日本人の生活に欠かせない表現である。現実世界の音や様態を表現できるだけでなく、自身の内に持つ感情・感覚をも表現できる。

このため、オノマトペは非常に数が多い。見出し語だけでも 4000 を超え (小野 2007)、さらに、新語が作られるため、掲載されない語も多い。

同様に、種類についても様々である。例えば、擬音語、擬態語、擬情語といった分類が挙げられる。

1.1. 擬音語・擬態語

擬音語は外界の音に由来する表現であり、物理的な音を言語で表現しようとした結果である。物理音由来のため、意思の疎通が図れなければ、その音を直接聞かせればよい。例えば、日本語学習者に『わんわん』が伝わらないなら、鳴き声を聞かせればよい。おそらく、『Bow-Wow』であると気づくだろう。

次の擬態語は物事の様態に由来する表現である。例えば、『ぶかぶか』等が挙げられ、擬音語同様、人間の知覚に訴えた意思の伝達が可能である。もし『ぶかぶか』が伝わらなければ、大きなズボンを履いて見せればよい。おそらく、『baggy』と判断しよう。

1.2. 擬情語～感情・感覚を表すオノマトペ

前節に述べたように、擬音語や擬態語は直接観測可能である。しかし、感情・感覚を表す擬情語については、知覚を通じて意思の疎通を図ることは難しい。個人の感情に根ざした表現だからである。例えば、痛みや痒みに伴う「むずむず」という単語の意味を簡単に伝えられるだろうか？ この説明は日本人にとっても容易ではないと考えられる。

1.3. オノマトペのクラスタリング

さらに「むずむず」は『何かしたくてたまらない』ことも示す。つまり、オノマトペは多義的である。

しかし、分類研究では精度等を求める手法上の制約から多義性を考慮せず、k-means, Kohonen の自己組織化マップ(2001)、階層的手法等のハードクラスタリング手法の採用も多く (浅賀ら 2007, Kurosawa et. al. 2010, 古宮ら 2011)、多義性を反映しているとは言い難い。そこで fuzzy c-means を用い、ソフトクラスタリングによる効果を検討する。

2. fuzzy c-means

本研究では、Bezdek (1981) による fuzzy c-means を用いる。様々な検討がなされ、派生アルゴリズムの展開も行われ、有効な手法と考えられている。

2.1. fuzzy c-means のアルゴリズム

n 個のデータを c 個のクラスタに分類するため、次の目的関数を考える。ここで帰属度 u_{ik} は $n \times c$ 行列、距離 d_{ik} は $d_{ik} = \|x_k - v_i\|^2$ を示す。 x_k は k 番目のデータを、 v_i は i 番目のクラスタ中心を表す。

$$J = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m d_{ik}$$

有効なクラスタを獲得 (J を最小化) するため、以下の 2 式の計算 (クラスタ中心の計算及び帰属度の計算) を繰り返し行う¹。また、パラメータ m ($m > 1$) は曖昧さを決定する。 m が 1 に近づくと曖昧さは減少し、k-means 法の結果と同様になる。

$$v_i = \frac{\sum_{k=1}^n (u_{ik})^m x_k}{\sum_{k=1}^n (u_{ik})^m}$$

$$u_{ik} = \left[\sum_{k=1}^n \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{1}{m-1}} \right]^{-1}$$

2.2. クラスタへの帰属確率

最終的に個々のデータの帰属度が確率で表現される。結果例を次に挙げる (表 1)。この例は、2 個のクラスタへの分類例～ $c = 2$ を指定～である。

このように複数のクラスタへの帰属が記述可能であり、多義のオノマトペにも対処可能と考えられる。

表 1 fuzzy c-means による結果例

		帰属度 (クラスタ)	
		First	Second
User	A	0.95	0.05
	B	0.65	0.35
	C	0.08	0.92
	D	0.15	0.85

¹ 導出については省略 (例えば、宮本(1999)等、参照のこと)。また、本研究は初期中心にデータ中のランダムな 1 点を選んだ。

3. 実験と考察

本研究は fuzzy c-means を利用し、オノマトペの分類、特に複数辞書の定義による分類を試みる。

3.1. 言語データ

3.1.1. 使用辞書

本研究はオノマトペと、複数の辞書に記載されたその定義文を用いる（論文末参照）。また、分類としては小野（2007）の分類を利用する。複数の分類項目への分類を許しているため、黒澤ら（2011）、Kurosawa et. al.（2010）が用いた、1項目への分類よりも有効であると考えられる。

特に、聞き手への説明が難しい感情・感覚に関連したオノマトペに着目する。そして、小野が『感情・感覚』に分類したオノマトペのうち、「あわてる・もがく・落ち着かない（39語）」「痛む（44）」「怒る・不機嫌・無愛想（40）」「思う・感じる（20）」「元気がない（30）」「ためらう・ひるむ（10）」「喜ぶ（11）」の7種（170語）を用い、獲得クラスタがこの分類と一致するか検討する。なお、「笑う」等の分類は「いひひ」等の擬音が多いため使用しないこととする。

3.1.2. 前処理

形態素解析器 MeCab（工藤）を使用し、自立語のみを処理の対象とする。ただし、1回しか出現しない語は削除する。また、「こと」などの形式名刺も使用しない。さらに、辞書の定義に頻出する語句、例えば、「ようす」「表す」等も使用しない。

上記手続きにより得られた自立語は879語であった。本研究はこの語から多次元ベクトルを構成した。

3.2. 予備実験

前項の言語材料を使用し、パラメータ決定のための予備実験を行った（表2）。表には、fuzzy c-meansの実行の際、パラメータ m の変化結果を載せている。

使用した言語材料では、 m の値が大きくなるにつれ、獲得クラスタ数が少なくなっていることがわかる。また、 m の値が大きいところでは、7種類のオノマトペ種のうち1種のしか検出していない。

一方、 m が小さく、また指定クラスタ数が大きい箇所では、獲得数も多く、多数のオノマトペ種の検出が可能であった。特に、 $m=1.2$ 、指定クラスタ数 $c=90$ の時、最大5種のオノマトペ種を検出した。

そこで、本研究は最大のオノマトペ種を検出できる値を採用し、以下の実験では $m=1.2$ に限定する。

3.3. 実験結果 1

実験結果を以下に示す。表3には精度が0.6以上になったクラスタの数を、表4には0.8以上のクラスタの数を示す。表中、確率値の上位5件（Top5）と上位10件（Top10）が記されている。

3.3.1. 獲得クラスタ数

指定クラスタ数が少ない場合、検出オノマトペ種が少なく、大部分が「痛む」「怒る・不機嫌・無愛想」に限定されていた。この種では分類されるべきオノマトペが多い（44個と40個）ためと考えられる。

一方、指定クラスタ数が大きくなると、多くのオノマトペ種が高い精度で検出可能となった。例えば、表3の100クラスタを指定した場合には、5種のオノマトペ種を検出できた。また、80、90など別のクラスタ数 c を指定した場合、別のオノマトペ種が検出された。しかし、先の考察と同様に「痛む」「怒る・不機嫌・無愛想」は多数検出される一方、「喜び（11個）」の検出回数は非常に少なかった。

3.3.2. 上位10件の精度

表4から、上位10件の大半は0または1クラスタを検出するだけである。本研究のデータでは、0.8以上の精度を求める課題は難しいことがわかる。

表2 パラメータ m の変化に伴う獲得クラスタ数（精度0.8以上のクラスタのみ）

		指定クラスタ数									
		10	20	30	40	50	60	70	80	90	100
m	1.1	3 (1)	3 (2)	5 (3)	9 (4)	8 (4)	10 (3)	16 (3)	20 (4)	19 (4)	13 (4)
	1.2	1 (1)	0 (0)	2 (1)	8 (3)	5 (3)	12 (3)	11 (4)	9 (4)	15 (5)	17 (4)
	1.3	4 (1)	3 (1)	0 (0)	1 (1)	5 (2)	6 (2)	6 (3)	5 (4)	15 (3)	14 (4)
	1.4	2 (1)	4 (1)	1 (1)	1 (1)	0 (0)	5 (2)	2 (2)	7 (2)	17 (3)	13 (4)
	1.5	5 (1)	6 (1)	3 (1)	1 (1)	1 (1)	2 (1)	0 (0)	9 (2)	5 (2)	11 (3)
	1.6	3 (1)	4 (1)	5 (1)	10 (1)	0 (0)	1 (1)	2 (2)	2 (2)	0 (0)	4 (1)
	1.7	2 (1)	6 (1)	4 (1)	2 (1)	2 (1)	3 (1)	1 (1)	2 (2)	10 (1)	5 (3)
	1.8	3 (1)	5 (1)	6 (1)	5 (1)	13 (1)	6 (1)	7 (2)	1 (1)	2 (1)	6 (1)
	1.9	3 (1)	6 (1)	8 (1)	3 (1)	5 (1)	5 (1)	1 (1)	2 (1)	1 (1)	2 (1)
	2.0	2 (1)	7 (1)	7 (1)	11 (1)	8 (1)	4 (2)	3 (1)	2 (1)	5 (1)	1 (1)

※ カッコ内は、全7種の内、何個のオノマトペ種を検出できたかを示す

3.3.3. 獲得オノマトペ種

前項に述べたように、オノマトペ種の獲得傾向には差がある。例として、90 クラスタ指定時 ($c=90$) の上位 5 位での精度について表に記す (表 5)。

表から、「元気がない」の精度 1.0 の項目では唯一のクラスタが獲得されているのに対し、「痛む」の精度 0.6 では、45 種のクラスタが獲得されている。

原因としては、もとの分類中のオノマトペ数に違いにあることに加え、データがスパースになっていることが考えられる。このことは、実際の獲得クラスタと、その帰属確率からわかる。「元気がない」の精度 1.0 の項目を一例として挙げる (表 6)。

表 6 オノマトペ種「元気がない」の一例

オノマトペ	とぼとぼ	すごく	よぼよぼ	ぐったり	げんなり
帰属確率	0.99411	0.01989	0.01469	0.01308	0.01202

第 1 位の 0.99411 に比べて、第 2 位の値が 0.01989 と非常に低い値になった。「とぼとぼ」と「すごく」の 2 語が、さほど空間上近接していないことを示す。

今回検討を行なっている fuzzy c-means は、距離 d_{ik} に依存した手法である。このため、スパースで次元数の大きい行列では、定義文中に共通の語があっても、その効果が消されると考えられる。

今回の課題では、「とぼとぼ」と「すごく」の 2 語が同種のクラスタに属す」とわかっている。しかし、実際の分類課題では、得られた値から分割の可否判断が必要とされる。そのとき、0.99411 と 0.01989 の 2 値が同一クラスタに属するという判断は難しい。

以上の結果、次元数の問題が影響を与えている可能性が示唆された。そこで、次元数の削減を試みる。

3.4. 実験結果 2

3.4.1. 前処理 2

次元数を減らすため、2 処理 (表記の揺れの統一、動詞・名詞等の統一) を追加する。

前者は、ある辞書では漢字表記の語が、別の辞書では仮名綴りとなることがあるため行う。また後者についても、ある辞書の「〇〇が動く」との記述が、別の辞書で「〇〇の動き」と、名詞・動詞変換が行われることもある。もちろん、複雑な言い換え規則や意味処理の適用も可能ではある。しかし、本研究では簡単な変換のみで次元数を減らすことを試みる。

以上の結果、次元数は 879 から 742 に減った。

3.4.2. 獲得クラスタ数

精度 0.6 以上については、表 3 の結果とほぼ変わらなかったため、0.8 以上のみ、分析を行う (表 8)。

3.3.2 で議論した表 4 と比べて明らかに異なる点は、指定クラスタ数 $c=100$ のところで 6 個のクラスタを獲得し、複数のオノマトペ種を検出している点である。表 4 では、どちらも 0 であったので、この点は大きな改善点であると言える。

3.4.3. 獲得オノマトペ種

次に、第 2 位の確率が変化したかについて検討する。一例として、上位 5 位の精度が 0.6 以上あった指定クラスタ数 90 のオノマトペ種の例を挙げる。

表 3 精度 0.6 以上のクラスタの数

		指定クラスタ数									
		10	20	30	40	50	60	70	80	90	100
Top	5	1 (1)	20 (2)	29 (2)	38 (3)	45 (3)	56 (3)	58 (5)	67 (5)	71 (5)	79 (5)
	10	1 (1)	1 (1)	1 (1)	8 (2)	9 (3)	13 (3)	11 (3)	2 (2)	14 (4)	17 (4)

※ カッコ内は何個のオノマトペ種を検出できたかを示す

表 4 精度 0.8 以上のクラスタの数

		指定クラスタ数									
		10	20	30	40	50	60	70	80	90	100
Top	5	1 (1)	0 (0)	2 (1)	8 (3)	5 (3)	12 (3)	11 (4)	9 (4)	15 (5)	17 (4)
	10	0 (0)	0 (0)	0 (0)	1 (1)	1 (1)	1 (1)	1 (1)	0 (0)	0 (0)	0 (0)

表 5 90 クラスタ指定時の確率と検出クラスタ数

	元気がない	痛む	怒る・不機嫌・無愛想	あわてる・もがく・落ち着かない	ためらう・ひるむ	痛む	元気がない	痛む	怒る・不機嫌・無愛想	計
精度	1.0	1.0	1.0	0.8	0.8	0.8	0.6	0.6	0.6	—
クラスタ数	1	4	1	3	1	5	1	45	10	71

表 7 例「あわてる・もがく・落ち着かない」

オノマトペ	うきうき	わくわく	そそくさ	ほくほく	ばたばた
帰属確率	0.48837	0.17413	0.02588	0.0147	0.01342

個の例では，“うきうき”と“わくわく”の 2 語が比較的近接に配置されている．このような例が多くなれば，適切な分割指標も作成でき，また多義的なオノマトペの記述にも使用可能となると考えられる．

3.4.4. 帰属確率第 2 位に関する分析

先の分析では第 1 位の語が孤立している可能性について述べた．ここでは帰属確率の第 2 位の語について分析を行う．変動をなくすため，指定数 $c = 90$ を 5 回実行し，階級毎に和を求めた（図 1）．

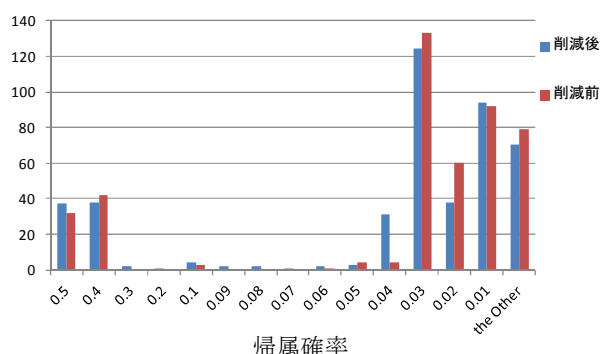


図 1 帰属確率毎ヒストグラム

0.04 や 0.3 においてわずかに上昇が見られる．しかし，当初予想したような効果ではなかった．この点は，依然として次元数が多いことを意味しており，さらなる削減が必要であると考えられる．

もう 1 点，妥当性指標（Gath and Geva 1989, Hashimoto ら 2009, 宮本 2009）についても述べておく．今回， W_{det} ， W_{ir} の 2 指標の計算を試みた．しかし，どちらも非常に小さな値となった．このことも，さらなる次元の削減の必要性を示唆している．

4. おわりに

本研究は複数の辞書の定義を利用し，感情・感覚に関連したオノマトペの分類を試みた．上位 5 件の精度を確認した結果，多くのクラスターで 6 割超の検出を行えていること，さらに次元数を減らした結果，上位 10 件の精度が 8 割を超えるクラスターを獲得できたことから，有効な検出が行えていると言える．

今後の課題としてはさらに次元数を減らし，各クラスター上位の帰属度の値を向上させ，多義的なオノマトペの記述精度を高めることである．また，その上で他の手法，例えば，Hofmann (1999) の pLSA と比較することも必要であろう．

謝辞

この研究の一部は，平成 22 及び 23 年度 広島市立大学特定研究費（一般研究）の補助を得ている．関係各位に感謝申し上げる．

参考文献

- 浅賀千里，ユスフ ムカルラマー，渡辺知恵美 (2007)．“オノマトペ用例辞典における用例を意味により分類するためのクラスタリング手法の諸検討”，日本データベース学会 Letters 6(2)．
- Bezdek, J. C. (1981), Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, NewYork.
- Gath, I. and Geva, A. B. (1989), “Unsupervised Optimal Fuzzy Clustering,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(7), pp. 773-781.
- Hofmann, T. (1999), “Probabilistic Latent Semantic Indexing,” in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,” pp.50-57.
- Hashimoto, W., Nakamura, T., and Miyamoto, S. (2009), “Comparison and Evaluation of Different Cluster Validity Measures Including Their Kernelization,” Journal of Advanced Computational Intelligence and Intelligent Informatics 13(3), pp. 204-209.
- Kohonen, T. (2001). “Self-Organizing Map, 3rd Edition.” 徳高平蔵，岸田悟，藤村喜久郎訳 (2005) “自己組織化マップ.” シュブリンガー・ジャパン．
- 古宮嘉那子，小谷善行 (2011), “階層型クラスタリングを利用した文脈によるオノマトペの分類”，NLP 若手の会 第 6 回シンポジウム．
- 工藤拓， “形態素解析器 MeCab.”，<http://chasen.org/~taku/software/mecab/>．
- 黒澤義明，竹澤寿幸 (2011). “自己組織化マップ SOM を用いた擬情語の分類比較～ 確率的潜在意味解析 pLSA による効果の検討 ～”，人工知能学会全国大会．
- Kurosawa, Y., Mera, K., and Takezawa, T. (2010). “Psychomime Classification and Visualization Using a Self-Organizing Map for Implementing Emotional Spoken Dialog System.” In Spoken Dialogue Systems Technology and Design, Wolfgang Minker, W. Lee, G. G., Nakamura, S., and Mariani, J. (eds), pp.107-134, Springer.
- 宮本定明 (1999), クラスタ分析入門，森北出版，POD 版，2010.
- 宮本定明 (2009), “ファジィクラスタリングの有用性について”，知能と情報，Vol. 21(6), pp.1008-1017.

定義文として利用した辞書

- 浅野鶴子 編 (1978), “擬音語・擬態語辞典”，角川書店．
- 阿刀田稔子，星野和子 (1998), “擬音語擬態語使い方辞典第 2 版”，創拓社．
- 飛田良文，浅田秀子 (2002), “現代擬音語擬態語用法辞典”，東京堂出版
- 松村明 編 (1995), “大辞林 第 2 版”，三省堂．
- 小野正弘 編 (2007), “日本語オノマトペ辞典”，小学館．
- 新村出 編 (1991), “広辞苑 第 4 版”，岩波書店．
- 山口仲美 編 (2003), “暮らしの言葉 擬音・擬態語辞典”，講談社．

表 8 精度 0.8 以上のクラスターの数（次元数削減後）

		指定クラスター数									
		10	20	30	40	50	60	70	80	90	100
Top	5	3 (1)	1 (1)	1 (1)	2 (2)	6 (4)	11 (3)	12 (4)	14 (4)	13 (4)	38 (4)
	10	0 (0)	1 (1)	0 (0)	0 (0)	2 (1)	2 (1)	0 (0)	1 (1)	17 (1)	6 (2)