自己組織化マップ SOM による
心情を表すオノマトペの意味分類と可視化

中村 沙織　　黒澤 義明　　竹澤 寿幸
広島市立大学 情報科学部　　広島市立大学大学院 情報科学研究科
{nakamura,kurosawa,takezawa}@nlp.its.hiroshima-cu.ac.jp

1.はじめに

オノマトペとは擬音語・擬態語の総称である。オノマトペは物事を的確に表現することが可能で、コミュニケーションを図る上で重要である。しかし、オノマトペが感覚的な表現であるため、語義が非常に曖昧であり、的確な意味を理解することが困難となっている。

加えて日本語にはオノマトペの種類や表現が多く存在し、辞書に見出し語として収録されている語だけでも約 4500 語も存在する(小野 2007)。また、オノマトペは新語が絶えず作られるため、辞書に載っていない語も多いと考えられる。

この様な理解困難さを解消し、新語への対応を容易にするため、本研究では Web から用例を抽出し、オノマトペを自動分類するシステムを提案する。そして、k-means 法でクラスタリングする手法(浅賀ら 2007)等の先行研究と比較を試みる。

2.自己組織化マップによるオノマトペの可視化

オノマトペは係っている動詞によって意味が大きく異なると考えられる。そこで本研究では、動詞項構造シソーラス(竹内 2008)を用いて、対象となるオノマトペと共起性の高い動詞を意味概念によってグループ化し、動詞の概念ごとにオノマトペを可視化する。この手段として、Kohonen(Kohonen 2001)による自己組織化マップ(Self-Organizing Map, SOM)を使用する。

SOM は、多次元ベクトルにより表されたデータを、その特徴を残し、他のデータとの相互関係を保ったまま、2 次元マップに写像することが出来る。すなわち、多次元のデータの関係を 2 次元平面上の距離として表し、視覚的に理解し易いと言う特徴を持っている。

2.1.自己組織化マップのアルゴリズム

SOM は二層からなる神経回路網モデルであり、教師なし学習を行う。入力層への特定の入力により、競合層の特定の領域が反応するような学習が行われる。入力層への入力ベクトル x は n 次元の広がりを持つベクトルであり、$x=\{x_1, x_2, \cdots, x_n\}$ のように表現される。また、競合層にはノードと呼ばれるユニットがあり、すべてのノードから、入力層との間に参照ベクトルと呼ばれるリンクが行われている。ここで、次式を満たす勝者ノード c の発見を試みる。次式は入力ベクトルに最も類似した参照ベクトルを持つノードを見つける操作と考えられる。

$$\forall i,\quad \|x-m_c\| \le \|x-m_i\|$$

上記の勝者ノードを発見したら、参照ベクトルを入力ベクトルに近づける操作を行い、さらに類似度を増すように学習させる。以下に、時間軸 を用いて表現した式を示す。

$\forall i \in N_c(t)$ を満たすとき、

$$m_i(t+1)=m_i(t)+h_{ci}(t)(x(t)-m_i(t))$$

それ以外のとき、 $m_i(t+1)=m_i(t)$

ある時間 t に使用した参照ベクトルを、入力ベクトルに似せて、次の時間 に参照ベクトルとして使用する、つまり、時間とともに、入力ベクトルに似ることになる。

ここで挙げた SOM は自然言語処理に適用され、分類の有効性が確認されている(cf. 黒澤ら 2008; 神崎ら 2007; 金 2003; 馬ら 2001)。

2.2.本研究の流れ

本研究では SOM による分類を行うための用例として Web コーパスを使用する。まず、オノマトペ、及びその後方一語に出現する動詞を取得する。後方一語の動詞に着目すると非常に語彙が多様となるため、動詞項構造シソーラス（竹内 2008）を用いて動詞を概念ごとに分類し、概念ごとに出現頻度を求める。そして、データ変換した上で自己組織化を行う。

2.2.1.使用コーパス及び対象語

本研究はコーパスとして Web 日本語 N グラム(工藤ら 2007)を使用する。

また、対象語としては心情を表すオノマトペ 228 語(Akita 2006)を用いる。そして片仮名表記、及び平仮名表記、さらに対象語の後方に「っ」、「ッ」、「っと」、「ット」を含む語を実験に用いる。

2.2.2.動詞の抽出

本研究ではコーパス上に出現したオノマトペの後方一語の動詞に着目する。特に、7-gram に記載された形態素列を合成した日本語文字列を使用する。この文字列を形態素解析器茶筌(松本 2000)により形態素解析し、対象語の直後に出現する形態素のうち動詞のみを収集する。なお、平仮名表記された動詞には解析誤りが多いため、平仮名表記の動詞は使用しないこととした。

2.2.3.動詞項構造シソーラス

動詞項構造シソーラスの内、本研究では特にフレームに着目をして分類を行う。ただし、1 つの動詞に対し複数のフレームが存在する動詞については、自動でフレームを決定することができないため今回は使用しないこととした。

2.2.4.データの変換

複数のベクトルの向きが同じで、極端に頻度に差があるとき、そのベクトルが示す語彙同士を異なる性質とみなす場合がある。よって本研究では動詞フレームの出現回数をオノマトペごとに出現率で示し、ベクトルとした。v_i の出現率を以下の式で計算する．動詞フレーム軸を $v_1, v_2, v_3, \cdots, v_n$ とする。

$$v_i\text{の出現率} = v_i \Big/ \sum_{i=1}^{n} v_i$$

3.実験と考察

3.1.実験方法

2 章で説明した手続きにより、このデータを元に自己組織化マップで分類を行った。データはオノマトペ 120 語、フレーム軸数 165 となった。なお、学習は 2 段階で行った。使用したパラメータは以下の通りである。結果を図 1 に示す。

・マップサイズ：64×48

・1: 学習回数 100,000，初期学習率係数 0.05

・2: 学習回数 1,000,000，初期学習率係数 0.01

3.2.実験結果

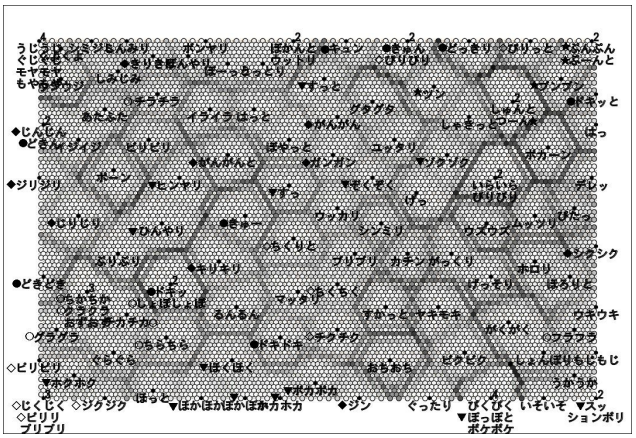

図 1　SOM による分類結果

図 1 の分類結果をさらに詳しく考察するため、対象語の分類を行った。対象語にはそれぞれ以下のように英語で意味が示されており(Akita 2006)、この説明文のうち特徴を示す単語に着目をした。

- dokidoki　‘feeling one’s heart *throb*bing,’
- hinyari　‘feeling pleasantly *cool*,’

なお、図 1, 2 の記号により、どのグループに属すかを示した。表 1 の分類を用いて新たに考察を加える。

表 1　考察用対象語の正解分類

特徴	記号	個数	着目した単語
目眩	○	9	eye, dizzy
鼓動	●	9	throb, heart jump
痛み	◆	10	sore, pain, pang,griping
臭い	★	5	smell
温度	▼	14	cool, chilly, cold, hot, warm, glowing
刺激	◇	9	pungent, skin
他		64	

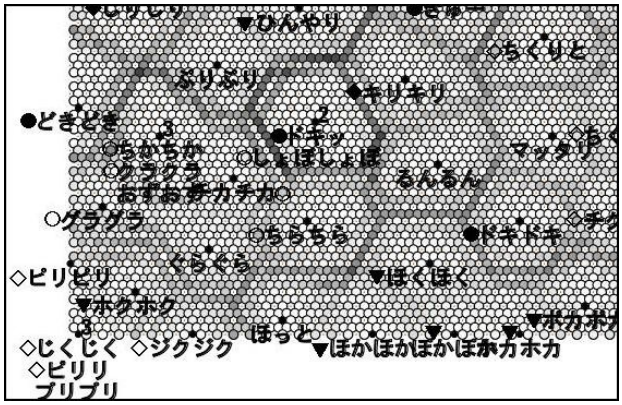

図 2　SOM による分類結果　左下 1/4

図 2 に SOM による分類結果の左下 1/4 の拡大図を示す。表 1 の目眩と温度が分類されていることがわかる。さらに温度のグループは寒暖が分かれてマップ化されている。すなわち、図 2 の右下から中央下にかけての暖（ポカポカ、ほくほくなど)、中央上周辺の寒（ひんやり、ヒンヤリなど）である。

3.3.考察

3.3.1.精度と再現率

表 1 のグループ 6 つ(目眩、鼓動、痛み、臭い、温度､刺激)に対してグループに属するオノマトペごとの軸のうち最もグループ全体に影響している軸をそれぞれ 3 つ選び精度・再現率を求めた。

ここで形態素解析誤りによる、データの誤差を除外するため、閾値を設定する。今回はベクトルごとの割合が 1/4、1/3、1/2 以下となる値の小さいベクトルの語彙を削除し、精度、及び再現率を求めた。

表 2　グループごとの精度・再現率

	目眩	鼓動	痛み	臭い	温度	刺激	目眩	鼓動	痛み	臭い	温度	刺激
	閾値なし						閾値1/4					
精度%	52.9	23.1	19.2	36.4	66.7	20.7	77.8	36.4	20.0	57.1	70.0	42.9
	9 / 17	6 / 26	5 / 26	4 / 11	8 / 12	6 / 29	7 / 9	4 / 11	4 / 20	4 / 7	7 / 10	6 / 14
再現率%	100.0	66.7	50.0	80.0	57.1	66.7	77.8	44.4	40.0	80.0	50.0	66.7
	9 / 9	6 / 9	5 / 10	4 / 5	8 / 14	6 / 9	7 / 9	4 / 9	4 / 10	4 / 5	7 / 14	6 / 9
	閾値1/3						閾値1/2					
精度%	87.5	44.4	21.1	57.1	70.0	46.2	85.7	66.7	25.0	80.0	70.0	60.0
	7 / 8	4 / 9	4 / 19	4 / 7	7 / 10	6 / 13	6 / 7	4 / 6	4 / 16	4 / 5	7 / 10	6 / 10
再現率%	77.8	44.4	40.0	80.0	50.0	66.7	66.7	44.4	40.0	80.0	50.0	66.7
	7 / 9	4 / 9	4 / 10	4 / 5	7 / 14	6 / 9	6 / 9	4 / 9	4 / 10	4 / 5	7 / 14	6 / 9

表 2 から、除外する対象を多くすると精度は上がる。しかし、再現率の低下が見られる。例えば、目眩のグループでは、閾値なしから 1/2 にすると精度が 52.9%から 85.7%に上がっているが､再現率は 100%から 66.7%に下がってしまう。このことから目的に応じて閾値の検討が必要であると考えられる。例えば、第二言語学習者の様なオノマトペについて知識がない場合、誤った知識とならないために精度が高いことが優先される。

3.3.2.SOM 分類の有効性

次にこれまでの研究と比較を行うため、対象データを他の分類手法で分類した。

まず、階層的クラスタリングとしてユークリッド距離を用いた最遠隣法で分類を行った結果を表 3 に示す。クラスタ数を 10 個に設定した。

表 3　最遠隣法の結果

クラスタ	1	2	3	4	5
個数	1	111	1	1	1
語彙	ウッカリ		ぞくぞく	うっとり	すっと
クラスタ	6	7	8	9	10
個数	1	1	1	1	1
語彙	がんがん	ガンガン	マッタリ	じりじり	しみじみ

表 3 より、最遠隣法では、1 つの語彙しか含まれないクラスタと、残りの語彙が集まったクラスタ(2)となった。

次に、非階層的クラスタリングとして浅賀ら(2007)が用いた k-means 法で分類を行った結果を以下に示す。k-means 法ではクラスタ数の設定により結果が大きく異なるため、考察のために用

いた表 1 のグループ 6 つ(目眩、鼓動、痛み、臭い、温度、刺激)に加えその他を分類するため、クラスタ数を 7 に設定した。

表 4　k-means 法の結果

クラスタ	1	2	3	4	5	6	7
個数	7	4	2	15	83	4	5
語彙	ほくほく ぽかぽか ジン	ぴりっと ぷんぷん ぷーんと	しゅんと つーん	うじうじ シミジミ ぽかんと	どっきり しみじみ チラチラ	キュン ぴりぴり ツン	あたふた どきん イジイジ

表 4 より、k-means 法では最遠隣法よりはバラつきがある。しかし、温度のグループの暖（ほくほく、ぽかぽかなど）はクラスタ(1)に分類されていたが、寒（ひんやり、ヒンヤリなど）はクラスタ(5)に含まれてしまい分類されなかった。

3.3.3 動詞項構造シソーラスの有効性

最後に動詞項構造シソーラスを用いず、対象語の後方の動詞をそのまま軸として SOM での分類を行った。ここでは、表 1 のグループに属するオノマトペごとの動詞の軸のうち、グループ全体に影響している軸の値を軸ごとに平均し、平均値が最大の軸を上位 3 つ決定し精度、及び再現率を求めた。平均値を用いて自動的に軸を決定したため、同じ手続きで本手法の動詞項構造シソーラスを用いた SOM での分類に対しても再び精度及び、再現率を求めた。

表 5　動詞項構造シソーラス使用　精度・再現率

	動詞項構造シソーラス 使用						動詞項構造シソーラス 不使用					
	目眩	鼓動	痛み	臭い	温度	刺激	目眩	鼓動	痛み	臭い	温度	刺激
精度 %	33.3	23.1	36.8	19.2	66.7	23.3	26.3	50.0	36.4	71.4	42.9	36.8
	8 / 24	6 / 26	7 / 19	5 / 26	8 / 12	7 / 30	5 / 19	4 / 8	8 / 22	5 / 7	9 / 21	7 / 19
再現率 %	88.9	66.7	70.0	100.0	57.1	77.8	55.6	44.4	80.0	100.0	64.3	77.8
	8 / 9	6 / 9	7 / 10	5 / 5	8 / 14	7 / 9	5 / 9	4 / 9	8 / 10	5 / 5	9 / 14	7 / 9

表 5 より、動詞項構造シソーラスを用いると目眩のグループでは精度、および再現率が高くなっているが、臭いのグループでは精度が大幅に低下している。このようにグループにより動詞項構造シソーラスの有効性が異なる結果となった。

4.おわりに

階層的・非階層的クラスタリングの両手法と比較した結果から、本手法での SOM による分類の有効性が確認された。

今後の課題としては、動詞項構造シソーラスによる分類では、1 つの動詞に対し複数のフレームが存在する多義語の分類が出来なかったため、グループごとでの有効性に差が生じたと考えられる。そのため、多義語への対処も検討の必要がある。

参考文献

Akita, K. (2006). “Embodied semantics of Japanese psychomimes.” KLS26: Proceedings of the Thirtieth Annual Meeting of Kansai Linguistic Society, pp. 45-55. 関西言語学会.

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 Vol.6, No.2.

金明哲 (2003). “自己組織化マップと助詞. 分布を用いた書き手の同定及びその特徴分析.” 計量国語学, pp.369-386, 計量国語学会.

神崎享子, 戸室宣子, 井佐原均 (2007). “自己組織化マップによる形容詞抽象概念の階層関係・類義関係の自動抽出.” 言語処理学会年次大会, pp.986-989.

Kohonen, T.(2001). “Self-Organizing Map, 3nd Edition.” 徳高平蔵, 岸田悟, 藤村喜久郎訳 (2005). “自己組織化マップ.” シュプリンガー・ジャパン.

工藤拓, 賀沢秀人 (2007). “Web 日本語 N グラム第 1 版.” 言語資源協会.

黒澤義明, 原章, 市村匠 (2008). “換喩検出を目的とした自己組織化マップ SOM による物体の形状マップ生成.” 言葉と認知のメカニズム, pp.353-374, ひつじ書房.

馬青, 神崎享子, 村田真樹, 内元清貴, 井佐原均 (2001). “日本語名詞の意味マップの自己組織化.” 情報処理学会論文誌, pp.2379-2391, 情報処理学会.

松本裕治 (2000). “形態素解析システム「茶筌」.” 情報処理, Vol.41, No.11, pp.1208-1214.

小野正弘 (2007). “擬音語・擬態語 4500 日本語オノマトペ辞典.” 小学館.

竹内孔一, 乾健太郎, 竹内奈央, 藤田篤 (2008). “意味の包含関係に基づく動詞項構造の細分類.” 言語処理学会年次大会, pp.1037-1040.