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

Examinations of Clustering Technique for Classifying Sentences by Meaning of Onomatopoeia on Online Onomatopoeia Example-based Dictionary

浅賀 千里♥　ユスフ ムカルラマー♦
渡辺 知恵美♠

Chisato ASAGA　Mukarramah YUSUF
Chiemi WATANABE

オノマトペとはいわゆる擬態語・擬音語のことである．事象を的確に表現でき，コミュニケーションを図る上で重要なものである．ところが，オノマトペは感覚的なものであるので外国人の日本語学習者がオノマトペの用例を習得するのは難しく，その学習に有効なのはオノマトペを含む文章を知ることだと言われている．そこで，我々は Web から数多くの新しい文章を抽出し，学習者に提示できるようなオノマトペ用例辞典の開発を進めている．

本辞典では現在，オノマトペの品詞的役割によって用例を分類し画面上に提示しているが，それでは様々な用例が混在し，オノマトペの意味が理解しにくい．そこで，学習者がより理解を深められるよう，本稿ではオノマトペの用例をオノマトペの意味ごとに分類し提示する手法を検討している．

Onomatopoeia which is imitative word expresses concrete phenomenon and plays a crucial part in communication. But it is difficult for Japanese learners who study Japanese onomatopoeia to understand its meanings and usages because onomatopoeia is sensuous. An effective method known to master onomatopoeia is to read a lot of sentences with onomatopoeia. There, we are developing an online onomatopoeia example-based dictionary which collects a lot of sentences with onomatopoeia from the Web.
The dictionary presents sentences which are classified according to word class of the onomatopoeia. However, each onomatopoeia has multiple meaning, and in the current version, sentences which have different meaning of onomatopoeia are mixed in the example list. There, we attempt to classify these examples of onomatopoeia by onomatopoeia meaning to present them to the Japanese learners in more understandable way.

♥ 学生会員　お茶の水女子大学大学院人間文化創成科学研究科博士前期課程　asaga@db.is.ocha.ac.jp
♦ 学生会員　お茶の水女子大学理学部情報科学科　mukarramah@db.is.ocha.ac.jp
♠ 正会員　お茶の水女子大学大学院人間文化創成科学研究科　chiemi@is.ocha.ac.jp

1. はじめに

オノマトペとは，いわゆる擬態語・擬音語のことである．具体的な事象を的確に表現できるのでコミュニケーション上重要なものであるが，オノマトペが感覚的な語であることや，外国語にオノマトペの対応語がないこと，１つのオノマトペが複数の意味を持つことなどから，日本語学習者にとってオノマトペの用例を習得するのは難しいといわれている．日本語学習者がオノマトペの意味・用法を理解するには，複数の用例として適切な文章からオノマトペが文中でどのように使われているのかを知ることが有効である．また，オノマトペは時代と共に意味が変わっていくことから常に新しい用例を得ることが重要となる．そこで，我々は Web から数多くのオノマトペの最新の用例を自動抽出し，日本語学習者に提示するようなオノマトペ用例辞典の開発を進めている．本辞典は Web から用例を抽出するため，適切ではない文章も抽出されてしまうので，語尾に付属語をつけると特定の品詞の役割を果たすというオノマトペの文法的性質を利用して用例を収集することにした．オノマトペに付属語をつけたものを見出し語として検索し用例を抽出する実験を行ったところ良い結果が得られたので，この方法を採用している[1][2]．

現在，本辞典は，抽出した用例をそのオノマトペの品詞的役割を元に分類し，画面に提示している．本稿では，複数の意味を持つオノマトペの用例をオノマトペの意味ごとに分類するために現在行っている諸検討の項目について報告する．「がりがり」を例にあげると，「氷をがりがり食べる．」や「がりがり勉強する．」，「がりがりの体．」など同じ「がりがり」でも全く違う意味を持つ．よって，このような異なる意味ごとに用例を分類し，提示する．具体的には，情報検索の手法を適用し，用例を文書ベクトルで表してクラスタリングする．また，適切な重み付けやオノマトペ辞書の効果的な利用について検討する．また，オノマトペとそれが係る用言に着目し，あるオノマトペが係る用言や，その用言がかかるオノマトペの関係を可視化することで，連鎖的に語彙の習得を支援するシステムについても検討する．

2. オノマトペ用例辞典

本研究で開発しているシステムはユーザがオンライン上で検索したオノマトペを含んでいる文章を Web から抽出し画面上に表示することで，日本語学習者にオノマトペの用例を提示するシステムである．用例の表示画面のイメージを図１に示す．学習者が用例を知りたいオノマトペを選択すると，そのオノマトペの用例が画面に一括表示される．

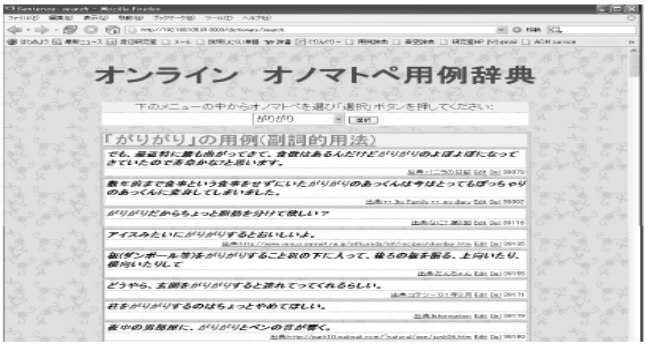

図１　インタフェースのイメージ
Fig.1　Image of Interface

2.1 システムの流れ

オノマトペリストからユーザが選択したオノマトペを含むWebページをYahoo!API[3]を用いて検索し自動取得し,その中からそのオノマトペを含んでいる文章を抽出する．検索を行う際，オノマトペに付属語を付けたものを見出し語として検索したページからも文章を抽出する．付属語については，2.2 節で述べる．抽出した文章を日本語係り受け解析器CaboCha[4]を用いてその文章の係り受け解析を行い，オノマトペが係る語句とその品詞を求める．抽出した文章と情報から用例としての適正を判定し，適正と判断したものをテーブルに格納する．用例提示用のWebサーバを設置し，データベースへの検索インタフェースを提供する[5]．

2.2 用例として適正な文章の抽出

本辞典はWebから文章を抽出しているため，用例として適正ではないものがいくつも検出される．その中には，オノマトペが複合名詞として使われているものやオノマトペ単体で使われているものが多い．そこで，しっかりとした文章になっている用例を抽出するために以下の手法を用例抽出手法に取り入れている．

1) Web から抽出した文章の中から，日本語係り受け解析器CaboCha を用いた時にオノマトペが何らかの語に係っていると判断された文章を抽出する．

2) オノマトペは語尾に付属語をつけることによって様々な品詞の役割を持つという文法的性質がある．例えば，「くるくる」に付属語「と」を付けると「くるくると」は「回る」や「転がる」などの動詞に係る副詞的役割をもつ．この性質を利用して，オノマトペの語尾に付属語をつけたものを見出し語としてYahoo!検索をし，文章を抽出する．

1)，2)の手法を用いることで，不適正な文章を除去し，用例として適切な文章が抽出できる割合を高めることができる．

3. オノマトペの意味による用例の分類

本辞典のインタフェースは，用例中のオノマトペが副詞的に使われているもの，連体詞的に使われているもの，複合名詞として使われているものというようにオノマトペの用法別に用例を分類して表示しているが，これでは，同じような意味で使われているオノマトペの用例が別用法の用例として表示されてしまう．例えば，副詞的用法の「がりがりのやせた雌犬で，そこそこ年はいってます.」と名詞的用法の「かなりがりがりな体.」という用例の「がりがり」は全て「ひどく痩せている」という意味であるが別々に表示される．これらの用例は同じグループとして表示した方が，より学習者は見やすくなるので，用例中のオノマトペの意味ごとに用例を分類して表示する手法を検討している．前述した「がりがり」の場合，既存の擬態語・擬音語の辞書[6]にある「がりがり」の意味は大きくわけて４つある．

1) かたい物を繰り返し引っ掻いたり，削ったり，噛み砕いたりした時などに発する音.
2) 引っ掻いたり，削ったり，噛み砕いたりした場合に，1)の音が発するようなかたさの様子.
3) ひどく痩せている様子.
4) 自分の欲望のために一途に物事に打ち込む様子
このように意味なよってオノマトペの用例を分類して提示することによって，学習者が理解しやすくなるようにする．

3.1 分類の基本手法

本研究では，用例分類の基本手法として情報検索手法を使用する．用例に出現する単語に重みをつけ，それを元に用例をそれぞれベクトル化し，ベクトルの距離から用例のクラスタリングを行う．手順は以下のようになっている．

あるオノマトペの用例がN個あったと仮定する．
1)それぞれの用例から主要な単語の語幹を抽出する．それ自体があまり意味を持たない単語の語幹は抽出しない．

2)それぞれから抽出した単語のIDFを計算する．
単語jのIDF=$\log(N/n_j)$
N:全用例数　n_j:単語jが含まれている用例数

3)それぞれの用例におけるその単語の重みを計算する．
用例iにおける単語jの重み $w_{ij}=t_{ij}$×単語jのIDF
t_{ij}:用例iに含まれている単語jの数

4)それぞれの用例について，単語jの方向に w_{ij} の大きさの用例ベクトルを作成する．

5)それぞれの用例ベクトルからクラスタリングを行う．

3.2 文書ベクトルの重みの調整

3.1 節の基本手法を用いた場合，用例中に含まれている単語はどれも同様に重要だとされて重みが計算されている．「私は昨日，がりがりと氷を削った.」という用例だと，基本手法のままでは「昨日」というあまり「がりがり」と関係のない単語と，「氷」や「削る」という「がりがり」に非常に関連ある単語の重要性が等しくなってしまう．オノマトペの用例から意味・用法を理解する際，オノマトペに関係する語が他の語に比べて特に重要であるので，オノマトペが係っている語（「削る」）・その語に係っている他の語(「氷」)の重みを強くする．

オノマトペの用例部分だけでは，オノマトペの意味がよく表せていないものもしばしば出てくるので，用例部分だけでなく，オノマトペの用例の周辺にある文章の全ての語に重みをつける．例えば，「猫が家の壁をかりかりと引っ掻いていた．犬もがりがりしていた．壁に引っ掻き傷がたくさんできた.」の場合，用例部分は「犬もがりがりしていた.」であるが，これだけだと何を表しているのかが明確に判断できない．そこで，周辺文章である「猫が家の壁をかりかりと引っ掻いていた.」,「壁に引っ掻き傷がたくさんできた.」に含まれるものに重みをつける．その場合，そのオノマトペの用例に位置的に近い文章の単語ほど重みは強く，用例から離れている文の単語の重みは弱くする．

3.3 クラスタリングの方法の検討

オノマトペの用例を分類する際に，どの手法がどのような利用に適切であるかを考察する．クラスタリング手法は，非階層的か階層的かに分かれ，更にそれぞれに対して，教師なし学習と教師あり学習に分かれる.3節冒頭に述べたように，辞書に定義されている意味ごとに分類したい場合には非階層的クラスタリングが適切であると考えられる．k-means法では辞書の定義の数を k に当てはめて分類するなどが考えられる．ただし，辞書では想定されないような使い方を感覚的に行えるのがオノマトペの特徴である．例えば，「がりがりの予算で働く.」という用例は辞書のどの定義にも当て

はまらない．そのため，辞書定義以外のクラスタも扱えるよう考慮すべきである．階層的なクラスタリングでは，非階層的なクラスタリングよりも細かい分類ができるようになる．例えば，「がりがり」では，3 節冒頭で述べたように大きくわけて 4 つの意味があるが，それぞれの意味に分類された用例の中で，更に細かく用例を分けることができる．1 つ目の「かたい物を繰り返し引っ掻いたり，削ったり，噛み砕いたりした時などに発する音」という意味で「がりがり」が使われている用例には，「犬が壁をがりがり引っ掻いている．」のようにペット等が家や壁，床を引っ掻いている様子を表す用例や，「がりがりと氷を食べる．」など，噛んだ時にがりがりと音がするような物を食べている様子を表す用例などがある．そこで，このような更に細かい意味等で階層的に分類することによって，きめ細やかな日本語の学習支援ができる．教師あり学習では，辞書や辞典的 Web サイトの用例が教師データとなるが，それぞれのオノマトペの用例が少ないため，効果的な学習が行えない．また，本の辞書から用例を抽出する場合，オノマトペの数自体が多いので用例数が膨大になり取得が困難である．このような理由から我々は教師なし学習を利用することを検討している．
これらのクラスタリング手法について，本辞典での使用方法や予想される結果等をより明確に考え，検討していく．

3.4 辞書を利用した分類手法

本研究では，既存の辞書や Web 上の辞典的なページを用いて用例の分類する．その手法を以下に示す．

1) k-means 法を使う場合，辞書にあるオノマトペのある意味の説明文をベクトル化し，それをシードとし，用例のクラスタリングを行う．
2) 辞書や辞典的なページにある，オノマトペのある意味に関する説明文の中からいくつかキーとなる単語を抽出し，その単語の重みを強くし，より用例が明確に分別できるようにする．例えば，3.1 節の「がりがり」の 1)の説明文の場合，
「かたい」，「物」，「引っ掻く」，「削る」，「噛み砕く」，「発する」，「音」が重要な単語となるのでその語の重みが強くなるようにする．

このように辞書を用いてクラスタリングを行うことを検討している．

3.5 予備実験

3.5.1 実験内容

用例のオノマトペの意味による非階層的なクラスタリングを行った際，それがどのくらい有効であるのか調べる目的で，以下の手法を適用して用例を意味により分類した場合，それぞれクラスタ内の用例がどのように分類されるのかを調べる実験を行った．使用したのは「がりがり」というオノマトペの Web きあら取得した用例 230 件である．実験で利用する「がりがり」の意味は 3 節冒頭で紹介した 4 つを利用する．今回は以下の 3 手法の比較を行った．

1) ランダムに取得した 4 つのシードを元に k-means 法でクラスタリングをする．どのシードにもあてはまらなかった用例はクラスタにランダムに挿入される．用例内の単語の重みはすべて同等である．

2)「がりがり」の 4 つの意味の説明文をベクトル化したものをシードにし，それを元に k-means 法でクラスタリングをする．どのシードにもあてはまらなかった用例はクラスタにランダムに挿入される．用例内で，オノマトペが係っている語の重みを通常の単語の重みの 4 倍にし，その語の係っているオノマトペ以外の語の重みを 2 倍にする．

3)「がりがり」の 4 つの意味の説明文をベクトル化したものをシードにして，まず，それぞれの用例を一番距離が近いシードに分類する．そして，どのシードのクラスタにもあてはまらなかったものがあった場合，クラスタを増やす．そして，クラスタ内の中心を決め，それぞれの用例を一番距離が近いシードに分類する．以上のことを繰り返すことで用例すべてをクラスタにわける．用例内で，語の重み調整は 2) と同様である．

3.5.2 実験結果

予備実験により，それぞれのクラスタ内の，意味(1)～(4)の意味でオノマトペが使われている用例の数をそれぞれ示す．

2) の手法では，はじめに意味(1)がシードとなっていたのはクラスタ 1,意味(2)がシードとなっていたのはクラスタ 2，意味(3)がシードとなっていたのはクラスタ 3，意味(4)がシードとなっていたのはクラスタ 4 である．よって，クラスタ 1 では意味(1)，クラスタ 2 では意味(2)，クラスタ 3 では意味(3)，クラスタ 4 では意味(4)の用例を中心に構成されているのが理想的である．

1) の結果を表 1, 2) の結果を表 2, 3) の結果を表 4 に示す．また，2) の結果の表 2 を元に再現率と適合率を計算したものを表 3 に示す．再現率は意味にあった用例の中でそのクラスタがどれだけの用例を含んでいるかという網羅性の指標であり，適合率はクラスタ内に得られた用例中にどれだけ意味に合った用例を含んでいるかという正確性の指標である．3) では，クラスタを増やしていった結果，最終的に 19 個のクラスタに分類された．

表 1　1)k-means 法(シード:ランダム)，重み調整なし
Table 1　Result of Experiment 1)

	意味(1)	意味(2)	意味(3)	意味(4)
クラスタ１	１７	8	4	0
クラスタ２	１６	1	１３	0
クラスタ３	１０	4	１０	5
クラスタ４	２７	6	１１	8

表 2　2)k-means 法(シード:辞書)，重み調整あり
Table 2　Result of Experient 2)

	意味(1)	意味(2)	意味(3)	意味(4)
クラスタ１	１２	3	１１	0
クラスタ２	２９	5	9	8
クラスタ３	２１	7	9	1
クラスタ４	8	3	１０	4

表 3　2)の再現率・適合率
Table 3 Recall and Precision of Experiment 2)

	再現率	適合率
A、クラスタ４	８/７０	８/２５
B、クラスタ３	７/１８	７/３８
C、クラスタ２	９/３９	９/５１
D、クラスタ２	８/１３	８/５１

表 4　3)k-means 法の拡張(シード:辞書), 重み調整あり
Table 4 Result of Experiment 3)

	意味(1)	意味(2)	意味(3)	意味(4)
クラスタ１	2	1	0	0
クラスタ２	8	2	6	0
クラスタ３	2	0	5	0
クラスタ４(C)(D)	１７	5	6	6
クラスタ５	1	0	1	0
クラスタ６	3	0	0	0
クラスタ７	2	5	0	0
クラスタ８	4	0	1	0
クラスタ９(B)	5	2	2	3
クラスタ１０	3	0	0	0
クラスタ１１	2	0	2	2
クラスタ１２	0	0	2	0
クラスタ１３	1	0	3	0
クラスタ１４	0	0	4	0
クラスタ１５	7	0	0	0
クラスタ１６	2	0	0	1
クラスタ１７	1	0	1	0
クラスタ１８	4	0	1	0
クラスタ１９(A)	3	0	0	1

3.5.3 考察

表 1 からわかるように，1)の手法では，1 つのクラスタにいろいろな意味の用例が混在してしまっている．よって，1)の手法をそのまま本辞典の用例の分類に適用することはできない．2)の手法では，辞書の説明文をシードとし，
オノマトペが係っている語と，その語に係る語の重みを強くしたので，1)の手法よりも良い結果が得られるのではないかと期待していたが，表 3 の再現率，適合率の低さからもわかるように，それぞれのクラスタは理想的な意味番号の用例で構成されていない．1)の結果と同様にいろいろな意味の用例が混在してしまった．その理由として，各ベクトルをクラスタに分類する際，どのクラスタにもあてはまらなかったベクトルを，ランダムに選んだクラスタに入れていることが挙げられる．実際，第 1 回目のクラスタリング処理で，どのクラスタにもあてはまらなかったものは 8 割以上であった．そのため，重みの調整や初期シードに関わらず，複数の意味の用例が混在してしまっている．3)の手法に関しては，クラスタ 6 内に意味(1)の用例のみが多く集まっているように，特定の意味を多く持つクラスタがたくさんある．他にはクラスタ 7 の意味(2)，8 の(1)，10 の(1)，12 の(3)，13 の(3)，14 の(3)，15 の(1)，18 の(1)，19 の(1)がある．辞書上で同じ定義に入るものが複数のクラスタに分割されている．それぞれのクラスタの中には，辞書の意味が更に細かく分類された意味ごとに用例が集まっていたことから，より階層的に用例を分類できるのではないかと考える．今後，これらの手法の改善や，他の手法についての検証を行い，より効率的に用例を分類できる手法を検討していく．

4. まとめと今後の課題

本稿では，オノマトペの用例を自動抽出し，それらを日本語学習者に提示するというオノマトペ用例辞典における用例の分類手法への取り組みについて述べた．
今後は，今回提案した手法を実際に実装し，検証していき，また検索インタフェースに機能を追加したりレイアウトを考えていくことでオノマトペ用例辞典が学習者にとってより使いやすいものになるようにしていきたい．

[謝辞]

本研究・開発に協力してくださった皆様に深謝する．

[文献]

[1] 浅賀千里, 渡辺知恵美.:``Web コーパスを用いたオノマトペ用例辞典の開発,"電子情報通信学会 第 18 回データ工学ワークショップ，B9-2 2007.
[2] 浅賀千里，渡辺知恵美.:``オノマトペのオンライン文例辞典の構築に向けて ,"第 25 回ことば工学研究会.
[3] ``YahooAPI,"http://developer.yahoo.co.jp/category/.
[4] 工藤拓 .:``CaboCha/ 南瓜 ,"http://chasen.org/~taku/software/cabocha/.
[5] George Chang , Marcus J.Healey , James A.M.McHugh and Jason T.L. Wang. :`` Web マイニング,"共立出版
[6] 山口仲美.:``擬音・擬態語辞典,"講談社.

浅賀 千里　Chisato ASAGA
お茶の水女子大学大学院人間文化創成科学研究科博士前期課程在学中．2007 お茶の水女子大学理学部情報科学科卒業．Web からの効率的な知識獲得の研究・開発に従事．情報処理学会学生会員．日本データベース学会学生会員．

ユスフ ムカルラマー　Mukarramah YUSUF
お茶の水女子大学理学部情報科学科在学中．2006 木更津工業高等専門学校卒業．Web からの効率的な知識獲得の研究・開発に従事．日本データベース学会学生会員．

渡辺 知恵美　Chiemi WATANABE
お茶の水女子大学大学院人間文化創成科学研究科講師．2003 お茶の水女子大学大学院人間文化研究科修了．博士(理学)．データベースシステムの研究・開発に従事．日本データベース学会正会員．