オンラインオノマトペ用例辞典「オノマトペディア」における効果的な用例分類法

浅賀 千里　(指導教員：渡辺 知恵美)

1　はじめに

オノマトペとは，擬態語・擬音語のことである．感覚的な語であるので，日本語学習者がオノマトペの意味・用法を習得するのは難しく，それらを習得するには，複数の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るため，最新の用例を知ることが重要である．そこで，本研究では，オノマトペの最新の用例を日本語学習者に提示するようなオノマトペ用例辞典「オノマトペディア」を開発している．

また，１つのオノマトペが複数の意味を持つ場合もあるため，用例をオノマトペの意味ごとにわけて表示するための用例分類手法を提案し，検証した．

2　オノマトペディア

オノマトペディアは Web から抽出したオノマトペの用例を表示し，オノマトペの意味・用法を学習してもらうシステムである．オノマトペ一覧から用例を見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表示される (図１)．

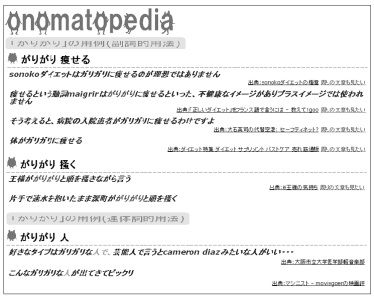

図 1: 用例表示画面のイメージ

2.1　システムの流れ

オノマトペを含む Web ページを Yahoo!API 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．抽出した文章の内，用例として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 [1].

2.2　用例として適正な文章の抽出

Web 上の文章には「ぽかぽか日記」などのブログのタイトルや「しっかり！」など主語や目的語が明確になっていないものなど，用例として不適切な文章がある．そこで，しっかりとした文章を抽出するために以下の手法を文章を抽出する際に取り入れている．

A　オノマトペを見出し語として検索して抽出した文章の内，CaboCha[4] を用いてオノマトペが動詞・名詞に係っている文章のみを用例とする．

B　オノマトペは語尾に付属語をつけることで特定の品詞の役割を持つという文法的性質があるので，オノマトペの語尾に付属語 (表 1) をつけたものを見出し語として検索をし，文章を抽出する [3].

C　まず，抽出法Ｂで収集した用例の表１で該当する部位を抽出する．「縄跳びで体ががりがりに痩せた.」が収集されたならば，主語の「体」と述語の「痩せる」を抽出し,「体が　がりがりと　痩せる」などオノマトペと組み合わせたものを見出し語として，文章として適正である可能性の高い用例を再収集する．なお，再検索は，オノマトペの係る語上位 30 語にオノマトペが係っている用例に対してのみ行う．

表 1: オノマトペと付属語の関係

Bの付属語	品詞	例	Cの抽出部
と，に	副詞	髪がさらさらと揺れる	主語，述語
な，の	連体詞	さらさらな髪	目的語
だ	形容動詞	髪がさらさらだ	主語
する	動詞	髪がさらさらする	主語

3　オノマトペの意味による用例の分類

オノマトペディアは現在，例えば,「身体ががりがりに痩せた」などオノマトペが副詞になっている用例と,「かなりがりがりの身体」などのオノマトペが連体詞になっている用例を，別のグループとして表示している．だが，この２つの用例の「がりがり」は共に「ひどく痩せている」という意味で使われているため，一緒に表示した方が学習者は「がりがり」の意味が理解しやすくなる．そこで，我々は用例中のオノマトペの意味ごとに用例を分類する手法を提案した．

3.1　文書ベクトルの作成

用例をそれぞれベクトル化し (文書ベクトル)，そのベクトルを元に用例のクラスタリングを行う．文書ベクトルは，単語を属性とし，その用例中でのその単語の出現回数を値としたベクトルを正規化したものとする．また，ベクトルを作成する際に，以下の調整を行う．まず，用例から意味・用法を理解する場合，オノマトペが係っている語・その語に係っている他の語が重要になるため，それらの語の重みを強くする．また，用例だけではオノマトペの意味を表せていない場合もあるので，用例周辺の文章の単語もベクトルに加える．

3.2　オノマトペの意味による用例の分類手法

オノマトペの意味により用例を分類するための予備実験として，2.2 節の抽出法Ａ，Ｂを用いて取得した用例を k-means 法を拡張した手法で分類した．クラスタ内の用例はオノマトペの係る語で強く結びついていた．異なる意味でオノマトペが使われている用例が混在してしまっているクラスタも複数あった．それらのクラスタには「する」や「なる」などの幅広く使われる単語にオノマトペが係っている用例が多く含まれていた．例えば,「体ががりがりになる」と「氷の表面ががりがりになる」では,「がりがり」意味は違うが共に「なる」に係っているので同じクラスタに入ってしまっていた．また,「がりがり」が「痩せる」に係っている用例と「体」に係っている用例は「がりがり」が同じ

意味のものが多いが別のクラスタに入ってしまっていた．そこで，オノマトペの係り受け関係を考慮したクラスタリング手法 [2] を提案した．また，更に効果的に分類を行うために，用例に含まれている単語の上位語を取り入れることにした．

なお，実験はすべて,「がりがり」の用例を,「がりがり」の意味ごとに分類する目的で行った．結果となる用例中の「がりがり」の意味は,「固い物を引っ掻いた時の音」,「ひどく痩せている様子」など７つにわけ，どれに該当するのかを手動で判断した．

3.3 係り受け関係を考慮したクラスタリング手法

3.2 節で述べた点を考慮し，係り受け関係を考慮したクラスタリング手法を提案した．主旨を以下に示す．

- 「がりがり」が「痩せる」に係っている副詞的用例,「がりがり」が「身体」に係っている連体詞的用例が別のクラスタになっているので，まず副詞的用法と連体詞的用法の用例を別々にし，それぞれのグループ固有の特徴を出す．
- オノマトペの係る語が同じ用例のオノマトペの意味はほぼ同じであるので，まずオノマトペの係る語で用例を分類する．
- 「する」,「なる」などの広く使われる語にオノマトペが係っている場合は，それぞれの用例に違う意味があるので他の用例とは別に扱う．

クラスタリング手順を以下に述べる．

(1) 副詞的用法，連体詞的用法の用例を別々にベクトル化する．

(2) オノマトペの係っている語で分類する．例えば，「がりがり」が「削る」に係る用例は同じクラスタになる．ただし，幅広く使われる「する」などに係る用例は，それぞれの用例を一つのクラスタとする．

(3) それぞれのクラスタに対する代表ベクトルを作成する．代表ベクトルは，そのクラスタ内の文書ベクトルを足して正規化したものとする．

(4) 副詞的用法の用例，連体詞的用法の用例全ての代表ベクトルを k-means 法でクラスタリングする

結果を図２に示す．表は左列が代表ベクトル，右列がクラスタ番号である．

描く	1	励む	2	言う	4	やせ細る	5	凸凹	8
でこぼこ	1	部屋	2	齧る	4	手ごたえ	5	書き込む	9
ひっかく	1	体型	3	かじる	4	やせる	6	思う	9
アイスバーン	1	痩せる	3	噛む	4	痩せっぽち	6	つて	9
引っ掻く	1	状態	3	体	4	身体	6	上	9
雪	1	氷	3	こする	4	人	6	狩る	9
擦る	1	斜面	3	立てる	4	胴	6	ガリ	10
書く	1	引っかく	3	粒状	4	掻く	6	挽く	10
戸	1			噛み砕く	4	かく	6	たてる	10
削る	1			食べる	4	使う	7		
凍る	1					方	7		
						行う	7		

図 2: 実験の結果

クラスタ１～7は「がりがり」が同じ意味で使われている用例が集まった．例えば，クラスタ１は固い物を引っ掻いたりした時の「がりがり」という音を意味する用例のクラスタである．クラスタ８～１０は，その中の単語に「がりがり」が係っている用例が少なかったことから，様々な意味の用例が集まってしまった．これにおいては，2.2 節の抽出法Ｃで用例の再収集を行うことで改善できると考えられる．

3.4 効果的な用例分類を行うための提案

オノマトペの意味による用例の分類を効果的にするため，以下の事項を取り入れる．

(1) 用例を抽出する際，2.2 節の抽出法Ｃを用い，用例の再検索を行うこと

(2) 文書ベクトルを作成する際，用例に含まれている単語の上位語を取り入れること

（１）は 3.3 節で述べた様に，オノマトペが係っている単語で用例件数が少ないものがあるため，それを改善するために，2.2 節の抽出法Ｃを用いて用例を集めることを考えた.（２）は用例内の単語の類義語や上位語を考慮するために行う．例えば,「がりがり」の用例である文章 A「犬が壁をがりがりと引っ掻いている」と文章 B「猫が壁をがりがりと掻いている」は,「動物が壁を削っている」という意味で類似しているが,「壁」しか共通していないため，基本手法では，この２つの用例の文書ベクトルの距離は近くならない．そこで，A の「引っ掻く」と B の「掻く」が類義語であることと A の「犬」と B「猫」の上位語は共に「動物」であることをベクトル作成時に考慮する．日本語の類義語・上位語辞典はコストが高いため，和英辞書と WordNet という英単語の上位語を取得できるアプリケーションを使用する．その手順を以下に示す．

1 用例中の単語を和英辞書を用いて英訳し，その英単語を文書ベクトルの属性に追加する．A の「引っ掻く」と B の「掻く」は英訳すると「scratch」となり文書ベクトルに共通して重みがつく．

2 用例中の単語を英訳したものを WordNet で解析し，その単語の上位語と，上位語の上位語を取得し文書ベクトルの属性に追加する．A の「犬 (dog)」と B の「猫 (cat)」の上位語は「animal」となり文書ベクトルに共通して重みがつく．

以上の手順でベクトルを作成しクラスタリングを行うことで，用例分類を更に効果的に行えると考える．

4 まとめと今後の課題

本稿では，オンラインオノマトペ用例辞典「オノマトペディア」におけるオノマトペの意味により用例の分類するための手法として，係り受け関係を考慮したクラスタリング手法を提案した．また，その分類をより効果的に行うための事項について述べた．

参考文献

[1] George Chang , Marcus J.Healey , James A.M.McHugh and Jason T.L. Wang. : “ Web マイニング,” 共立出版, 197p.

[2] 浅賀　千里，Yusuf Mukarramah，渡辺　知恵美 .: “オンラインオノマトペ用例辞典「オノマトペディア」における用例を意味により分類するための係り受け関係を考慮したクラスタリング手法 ,” DBWS2008.

[3] 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,” DEWS2007.

[4] 工藤 拓.: “CaboCha/南瓜,” http://chasen.org/ taku/software/cabocha/.