Web上のオノマトペの用例を共起単語で絞り込む用例抽出法

森田 一匡
山梨大学大学院
医学工学総合教育部
g08mk029@yamanashi.ac.jp

鈴木 良弥
山梨大学
工学部 コンピュータ・メディア工学科
ysuzuki@yamanashi.ac.jp

1. はじめに

オノマトペとは，「ワンワン」「さらさら」「スッキリ」などの擬音語や擬態語の総称のことである．日本語はオノマトペが豊富な言語[4]であるといわれており，日本人の感覚や感情を豊かに表現する言葉として日常生活の中でよく使われている．しかし，オノマトペは感覚的であるため，日本語を母語としない日本語学習者には意味・用法を理解することが難しい．そこで，日本語学習者が用例からオノマトペがどのように使われているかを学習することで意味・用法の理解に繋がると考えられている．

このような背景から，オノマトペを学習する日本語学習者の学習の補助を目的として，用例を学習するためにWeb上からオノマトペの用例を抽出する研究[1]や，類似した意味を持つオノマトペを学習するための辞書やシソーラスの作成といった研究，オノマトペの自動分類に関する研究[2]などが行われている．

現在，用例から文中でのオノマトペの使われ方を学習するシステム[7]がある．そのシステムでは「オノマトペ＋付属語(と，な，の，だ，する)」の文を適切な用例として抽出している．しかし，付属語がない文においても用例として適切なものが多いため，これらの文も抽出できることが望ましい．そこで，本研究では，Web上からより多くの用例を獲得することを目的として，付属語の有無に関わらず適切な用例を抽出する手法を提案する．まず，Web上から付属語がある文を用例として抽出し，そこから名詞・動詞の単語を抽出する．そして，付属語がない文の抽出条件として共起単語の情報を用いて絞り込む手法である．

検証実験では，goo版オノマトペディア[8]から選択した58語のオノマトペを用いて用例の抽出を行い，手法の有効性を検証した．また，付属語を用いていない文が，抽出された用例にどの程度含まれているかも検証した．

2. システムの流れ

本研究で作成したシステムの流れを図1に示す．このシステムは「Webページ取得部」，「用例候補文・共起単語取得部」，「用例取得部」の3つから構成される．

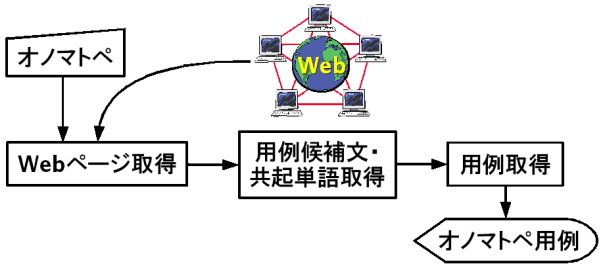

図1: システムの流れ

2.1 Webページ取得

本研究ではコーパスとしてWebページを用いる．大規模コーパスからのオノマトペの用例の抽出に関連する研究には浅賀ら[1]の研究がある．浅賀らはWebをコーパスとし，擬声語の文法的性質を取り入れることにより，質の良い用例を抽出することに成功している．文法的性質とは，擬声語の語尾に付属語（と，な，の，だ，する）を付加することにより様々な品詞（と/副詞，な/連体詞，の/連体詞，だ/形容動詞，する/動詞）の役割を果たす擬声語の特性である．本研究はまず，入力したオノマトペについて「オノマトペ＋付属語(と，な，の，だ，する)」を含むWebページ各50ページ分と，「オノマトペ」を含むWebページ50ページ分をYahoo!API[6]を用いて検索し，それぞれ取得する．また，オノマトペは基本的にはひらがなかカタカナのどちらか一方の表記で書かれているが，ひらがな⇔カタカナとなっても意味が変わるわけではない．そこで，ひらがな・カタカナ両方の表記で検索を行う．つまり取得するページは付属語あり（オノマトペ＋付属語）500ページ分と付属語なし（オノマトペのみ）100ページ分である．以上を表1にまとめて示す．

表 1: 取得 Web ページ数

	オノマトペ＋付属語	オノマトペのみ
ひらがな	250	50
カタカナ	250	50
合計	500	100

2.2 用例候補文・共起単語取得

取得した付属語ありの Web ページから用例候補文を取得し，その用例候補文から共起単語を取得する．

2.2.1 用例候補文抽出

取得した付属語ありの Web ページと付属語なしの Web ページからオノマトペが含まれている文を以下の定義に基づいてそれぞれ抽出を行う．

1. 付属語あり

   0 文字以上の漢字・ひらがな・カタカナ ＋オノマトペ ＋付属語 ＋1 文字以上の漢字・ひらがな・カタカナ

2. 付属語なし

   0 文字以上の漢字・ひらがな・カタカナ ＋オノマトペ ＋付属語以外のもの ＋1 文字以上の漢字・ひらがな・カタカナ

しかし，以下の定義に当てはまる文は抽出しない．

1. ひらがなとカタカナのみの文
2. すでに同一文が用例候補文中に存在する文
3. 用例候補文中の文の部分文字列である文

ここで抽出された文を用例候補文とする．

2.2.2 共起単語抽出

付属語ありの用例候補文を形態素解析システム MeCab[5]を用いて形態素解析を行う．抽出に使用する品詞の詳細を表 2 に示す．表 2 に示す品詞として解析される単語の基本形を全て抽出し，出現回数を求めておく．ただし，「ある」「する」のようなひらがな 2 字からなる動詞はオノマトペとの関連を判別しにくいと考えられるので抽出しない．また，基本形が存在しないものとオノマトペ自身も抽出しない．

表 2: 共起単語に使用する品詞一覧

品詞	品詞詳細
名詞	一般，固有名詞-一般，固有名詞-組織，固有名詞-地域，サ変接続，ナイ形容動詞語幹，形容動詞語幹
動詞	自立

2.3 用例取得

取得した共起単語を用いて付属語あり・なし両方の用例候補文から用例を抽出する．ここで，共起単語は出現回数が 10 回以上のものを使用する．ただし，10 回以上出現するものが 5 つ未満の場合は上位 5 つまでを使用する．(以下，共起単語とは用例抽出に用いる上記の共起単語のことを指すものとする．)

共起単語を用いて抽出する用例の定義は『出現回数第 1 位の単語を必ず含み，他の共起単語を最も多く含むもの』とする．抽出は付属語ありの用例候補文と付属語なしの用例候補文の両方から行い，定義に当てはまる全ての文を用例として抽出する．

3. 検証実験

作成したシステムを用いて共起単語抽出と用例抽出の精度を検証する実験を行った．実験には以下の 58 語のオノマトペを用いた．

「あせあせ」,「いじいじ」,「うきうき」,「かさかさ」,「かちかち」,「かばかば」,「からから」,「がりがり」,「ぎざぎざ」,「きちきち」,「きらきら」,「きりきり」,「くしゃくしゃ」,「くすくす」,「ぐだぐだ」,「けたけた」,「げたげた」,「ごーごー」,「ごくごく」,「こそこそ」,「ざあざあ」,「さくさく」,「さらさら」,「しくしく」,「しとしと」,「しゃきしゃき」,「じゃらんじゃらん」,「すかすか」,「すやすや」,「たじたじ」,「だらだら」,「ちかちか」,「ちくちく」,「ちらちら」,「つかつか」,「てかてか」,「どきどき」,「とろとろ」,「なでなで」,「にやにや」,「のろのろ」,「ぱかぱか」,「ばたばた」,「ぴかぴか」,「びしょびしょ」,「ひそひそ」,「ひゅるひゅる」,「ひゅんひゅん」,「ぶらぶら」,「ぷんぷん」,「へとへと」,「ほかほか」,「ぼちぼち」,「まじまじ」,「みしみし」,「むかむか」,「もこもこ」,「わくわく」

3.1 共起単語抽出の実験結果

本システムを使用し，抽出した共起単語の出現回数 1 位のものが適切であるかを調べる実験を行った．人手により，オノマトペとの関連が連想できるものを適切，関連が連想しにくい・できないものを不適切として分類を行った．また，出現回数第 1 位の単語が適切なものでない場合は上位 5 つまでに適切なものが出現しているかを調べた．実験の結果を図 2 に示す．

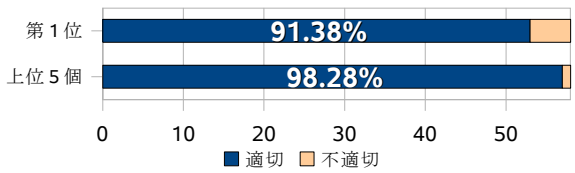

図 2: 共起単語抽出実験の結果

実験の結果，58 語中「あせあせ」,「がりがり」,「きちきち」,「ばたばた」,「もこもこ」の 5 語に

ついて第1位の単語が適切なものではなかった(8.62%). また，「もこもこ」については，上位5個以内にも適切な単語が出現しなかった(1.72%). 結果を表3，表4に示す.

表3: 出現回数第1位の単語が不適切なもの

オノマトペ	共起単語
あせあせ	あせる
がりがり	思う
きちきち	思う
ばたばた	炉端
もこもこ	カレー

表4: 上位5個全てが不適切だったもの

オノマトペ	共起単語	出現回数
もこもこ	カレー	24
	思う	15
	情報	11
	作品	9
	見る	9

3.2 用例抽出の実験結果

共起単語と同様に本システムを使用し，共起単語を用いて付属語ありの用例候補文と，付属語なしの用例候補文から用例を抽出し，抽出した用例が適切なものであるか調べる実験を行い，以下の4つのカテゴリに分類した.

1. 『適切』･･･ 適切なもの
2. 『不完全』･･･ オノマトペの使い方は正しいが，文として不完全(前後が欠けている)もの
3. 『特殊』･･･ オノマトペが一般的な使い方をされていないもの
4. 『不適切』･･･ 上記1~3のどれにも当てはまらないもの

実験の結果を図3，表5，表6に示す．表6は下線付きの単語がオノマトペ，斜体の単語が共起単語である．また，抽出された用例のうち付属語なしの用例候補文から抽出されたものがどの程度出現したかについても調べた.

実験の結果，58語中56語から用例を抽出した．用例は56語中から合計で208文抽出され，そのうち，付属語ありの用例候補文からは176文(84.6%)抽出され，付属語なしの用例候補文からは32文(15.4%)が 抽出された.

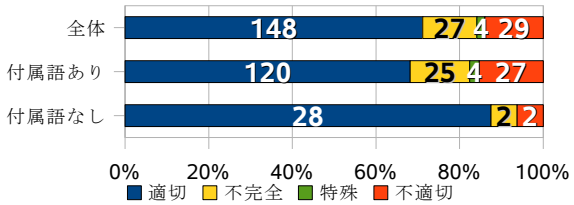

図3: 用例抽出実験の結果

表5: 各カテゴリの割合

カテゴリ	全体	付属語あり	付属語なし
適切	71.15%	68.18%	87.50%
不完全	12.99%	14.21%	6.25%
特殊	1.92%	2.27%	0.00%
不適切	13.94%	15.34%	6.25%

4. 考察

4.1 共起単語

図2，表7より，出現回数が多い共起単語はオノマトペと関連の強いことが証明できた．これは本研究で提案する共起単語を用いて用例を抽出する手法の有効性を示せたといえる．特に出現回数が30回以上の単語についてはオノマトペとの関連が非常に強いといえる.

4.2 用例

図3に示す通り，全体で71.15%の適切な用例を抽出することができた．付属語なしのものからは87.50%の適切な用例を抽出できた．これは付属語ありの用例候補文から適切な共起単語を抽出できていることによるものである．ただし，全体に含まれる付属語なしの用例の割合が少なかった．これは実験に使用しているWebページ数が付属語ありのページ計500ページ分に対し，付属語なしのページ計100ページ分であるためだと考えられる．また，付属語なしの検索結果には付属語ありの文も含まれてしまうため，用例候補文抽出時に必然的に付属語なしの文が少なくなってしまうことも考えられる．理論上は付属語なしの32文を5倍すると160文となり，付属語ありの176文と比率はほとんど変わらなくなる．ただし，再度実験をする必要がある.

5. まとめと今後の課題

本研究では，Web上からより多くのオノマトペ

の用例を獲得することを目的として，付属語の有無に関わらず適切な用例を抽出する手法を提案した．Web 上から付属語を用いて用例候補となる文を抽出し，共起単語を取得した．この共起単語の出現回数が上位のものを用いて付属語あり・なし両方の用例候補文から用例を抽出することで，付属語がない文からも適切な用例を抽出することができ，本手法の有効性を確認した．よって，用例の総数を増やすことができることを証明した．

今後はより高い割合で適切な用例を抽出できるように手法を見直すことを考えている．例えば，現在は共起単語を用いて抽出する用例の定義は『出現回数第 1 位の単語を必ず含み，他の共起単語を最も多く含むもの』としているが，この場合だと表 7 に示す単語で例を示すと，オノマトペ「さらさら」の共起単語「髪」を含む用例は抽出されない．そこで，出現回数第 1 位の単語に拘らず，『共起単語を多く含むもの』を優先することや，『名詞と動詞の共起単語を含むもの』を優先することなどを考えている．

また，オノマトペと共起単語の関連はオノマトペの分類にも応用できるのではないかと考えている．

参考文献

[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, 2007.

[2] 市岡健一,福本文代.: ”Web 上から取得した共起頻度と音象徴によるオノマトペの自動分類” 電子情報通信学会論文誌. D, 情報・システム 92(3) (447) pp.428-438, 2009.

[3] 三上京子.: ”上級教材に見られるオノマトペ-統語的特徴の分析と指導の観点- ”, 早稲田日本語教育研究第 2 号, pp.193-209, 2003.

[4] 国立国語研究所.: “擬音語・擬態語 -日本語を楽しもう！” http://dbms.kokken.go.jp/nknet/Onomatope/

[5] 高速形態素解析システム MeCab http://mecab.sourceforge.net/

[6] Yahoo!デベロッパーネットワーク http://developer.yahoo.co.jp/

[7] Onomatopedia http://www.onomatopedia.net/dictionary

[8] オノマトペディア - goo 辞書 http://dictionary.goo.ne.jp/onomatopedia/

表 6: 抽出された用例の分類例

カテゴリ	抽出した用例
適切	「*梅雨*のようにしとしと と長く*降り続き*，なかなか止まない*雨*」， 「毎日の*健康*は*血液*をサラサラ な*状態*に保つことです」， 「*横*では子供達がスヤスヤ と*寝*ていました」 「*家*全体がみしみし と*音*を*立て*て揺れ，立っていられないほどです」，
不完全	「日目の朝は肌寒く，*梅雨*のようなシトシト する*雨*が*降っ*ていました」， 「の*日常生活*とかをグダグダ と*書き*綴る」，「*音*はごーごー とうるさいものの」，
特殊	「太陽がけたけた *笑っ*ている」，「脳がケタケタ *笑っ*ていた」，
不適切	「*炉端*ばたばた の店舗情報」，「*カレー*のモコモコ のお店*情報*」， 「はてブがこそこそ だと*思っ*ている*人*」，

表 7: 30 回以上出現した共起単語

オノマトペ	共起単語	出現回数	オノマトペ	共起単語	出現回数	オノマトペ	共起単語	出現回数	オノマトペ	共起単語	出現回数
さらさら	血液 髪 流れる	143 69 36	かさかさ	肌 乾燥 皮膚	70 45 31	しとしと	雨 降る	161 31	にやにや	見る 笑う	41 30
						ざあざあ	雨	127	ぷんぷん	匂い	45
ごくごく	飲む 水	79 32	きりきり	胃 痛む	76 34	ちらちら	雪 見る	57 31	てかてか	顔	35
									たじたじ	質問	30
みしみし	音	50	けたけた	笑う	53	なでなで	頭	40			
からから	音	41	くすくす	笑う	51	すやすや	寝る	36			
しゃきしゃき	食	49	げたげた	笑う	42	しくしく	泣く	34			