造語の過程に基づく派生オノマトペの抽出

吉成 祐人　　藤井 敦
東京工業大学大学院情報理工学研究科計算工学専攻

1 はじめに

近年，不特定多数の人間がインターネット上で自由に情報を発信できるようになり，新しい語が生まれる機会が多くなった．擬音語や擬態語の総称であるオノマトペは，短い単語列で感性を表現できる便利さから，新しく造語される頻度が高い．造語された新オノマトペは自然言語処理に用いる辞書に載っていないため，解析を行う際に問題となる．そのため，新オノマトペをコーパスから自動的に抽出するための研究が行われている．

奥村ら [1] は，コーパスからオノマトペを抽出し，用例とともにオノマトペの概念辞書を自動構築する手法を提案した．具体的には，オノマトペに現れやすい音韻パターンを持つオノマトペを抽出の対象としている．

中島ら [2] は，既存のオノマトペから派生してオノマトペが造語される [3] 過程に着目して，コーパスから複合オノマトペを抽出する手法を提案した．具体的には，「さらさら」や「つやつや」のような,「ABAB」型の音韻パターンを持つ既存オノマトペを組み合わせて造語された「さらつや」のような複合オノマトペを抽出の対象としている．

奥村らは，新オノマトペが既存オノマトペから派生して造語されるという性質に着目しておらず，中島らは，2 つの「ABAB」型オノマトペを組み合わせた複合オノマトペしか抽出の対象としていない．

本研究は，新オノマトペが既存オノマトペから派生して造語されるという性質に着目し，複合オノマトペ以外の派生オノマトペを抽出する手法を提案する．具体的には,「ねちゃっ」を変形して造語された「ねっちゃり」のような，単体の既存オノマトペを変形して造語される派生オノマトペを抽出の対象とする．

2 派生オノマトペの抽出手法

2.1 概要

本研究で提案する派生オノマトペの抽出手法を図 1 に示す．本研究は，派生オノマトペの候補語を生成し，コーパスから既存オノマトペと候補語の用例文を収集した後に，オノマトペと判定された候補語を抽出する．

まず，音韻パターンに基づいて候補語の集合を機械的に生成する．次に，既存オノマトペと候補語の間で表記類似度と文脈類似度の特徴量を計算し，2 つの特徴量から「派生らしさ」のスコアを求める．オノマトペに後接しやすい文字列を伴う候補語の出現頻度から「オノマトペらしさ」のスコアを求める．2 つのスコアを統合した「派生オノマトペらしさ」に基づいて候補語を降順に整列し，順位が閾値以上である候補語を派生オノマトペとして抽出する．

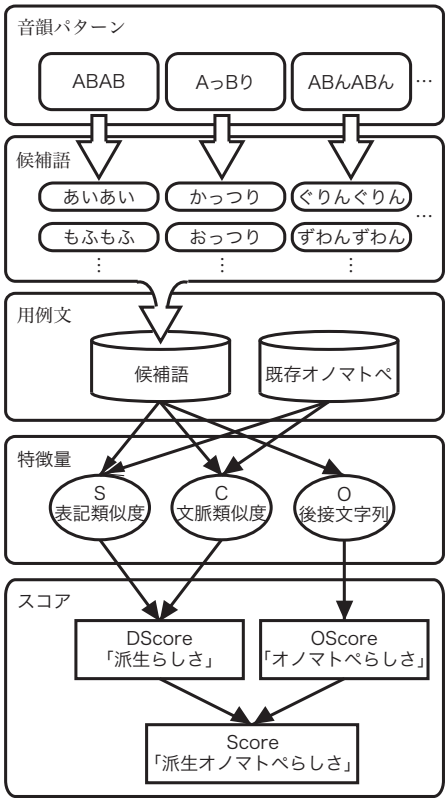

図 1: 派生オノマトペ抽出手法の概要

2.2 候補語の生成

奥村らと同様に音韻パターンに基づいて候補語の集合を生成する．音韻パターンには,「さくさく」のような「ABAB」型,「さっくり」のような「AっB り」型,「ばらんばらん」のような「AB ん AB ん」型が存在する．音韻パターンの A および B には 1 文字の平仮名（「あ」「い」「う」 …「が」「ぎ」「ぐ」…）と，拗音をもった 2 文字で 1 音節となる音節（「きゃ」「きゅ」「きょ」…）を当てはめる．ただし，B には撥音（「ん」），促音（「っ」），長音（「ー」）も当てはめる．A と B には異なる音節を

当てはめ，日本語として正しくない文字列（「っっ」「んん」「っー」）を含む語は生成しない．平仮名のみで構成される候補語だけでなく，カタカナのみで構成される候補語も生成する．

2.3 表記の類似度

派生オノマトペは派生元の既存オノマトペと表記が類似する傾向にある．そこで，派生オノマトペの候補語と既存オノマトペの間で表記の類似度を計算する．

具体的には，DP マッチングを用いて，編集距離を計算する．編集距離は，2 つの文字列を照合するために片方の文字列に対して行う「挿入」,「置換」,「削除」の操作回数を距離とし，計算される．ただし，本来の編集距離とは異なり，本研究では全ての操作を距離 1 とはせずに，派生が起こりやすい操作の距離を短くする．清音，濁音，半濁音の入れ替え（「き」→「ぎ」），拗音の挿入，置換，削除（「き」→「きゃ」），促音の挿入，置換，削除（「き」→「きっ」）を特殊操作として距離を 0.5 とし，それ以外の操作は距離を 1 とする．一つの音節に対して複数の特殊操作が適用される際（「き」→「ぎゃ」）は，式 (1) を用いて距離 D を計算する．

$$D = 1 - \sum_{i=1}^{n}(1 - d_i) \tag{1}$$

n は適用される特殊操作の数であり，d_i は i 番目に適用される特殊操作単体の距離である．

式 (2) を用いて候補語 w と既存オノマトペ o の表記類似度を計算し，表記類似度の特徴量 $S(w,o)$ とする．

$$S(w,o) = 1 - \frac{w \text{ と } o \text{ の編集距離}}{w \text{ と } o \text{ で音韻数が多い語の音韻数}} \tag{2}$$

$S(w,o)$ は，計算対象の語を構成する音韻数が多いほど，操作による影響が小さくなる．

2.4 文脈の類似度

中島らは，複合オノマトペは複合元の「ABAB」型オノマトペと意味的に類似している可能性が高く，似たような文脈で用いられると仮定した．そこで，複合オノマトペと複合元オノマトペの間で文脈類似度を計算した．

同様に，派生オノマトペは派生元の既存オノマトペと意味的に類似している可能性が高く，似たような文脈で用いられることが予想される．そこで，候補語と既存オノマトペの間で文脈類似度を計算して，文脈類似度の特徴量とする．中島らと同様に，候補語と既存オノマトペを共起語の重みを要素とするベクトルで表す．

まず，コーパスに対して MeCab で形態素解析を行う．候補語とその共起語，および既存オノマトペとその共起語の組み合わせに対して，それぞれ自己相互情報量（PMI）を求める．ある語 x とその共起語 c の PMI を式 (3) で計算する．

$$PMI(x,c) = \log \frac{p(x,c)}{p(x)p(c)} \tag{3}$$

$p(x,c)$ は x と c が共起する確率であり，$p(x)$ と $p(c)$ はそれぞれ x と c の出現確率である．

候補語 w とオノマトペ o のベクトルからコサイン類似度を計算して，文脈類似度の特徴量 $C(w,o)$ とする．

表 1: 奥村らがオノマトペ判定に用いた後接文字列

品詞	後接文字列
サ変名詞	する，した，して
タル形容詞	たる
ナ形容詞	だ，な，に
ナノ形容詞	の
名詞	が，は，を
副詞	と

2.5 「派生らしさ」のスコア

候補語 w と既存オノマトペ o に対して，表記類似度の特徴量 $S(w,o)$ と文脈類似度の特徴量 $C(w,o)$ を用いて，「派生らしさ」のスコア $DScore(w,n)$ を計算する．予備実験を行った結果，$S(w,o)$ は $C(w,o)$ よりスコアが上位に集中しやすい傾向があった．そのため，$S(w,o)$ と $C(w,o)$ を同等に扱うために，式 (4) を用いて $S(w,o)$ の値を調整し，調整後の $S'(w,o)$ を表記類似度の特徴量として用いる．

$$S'(w,o) = e^{a \times S(w,o)} - 1 \tag{4}$$

a はグラフの傾きを制御する定数であり，本研究では 3 に設定する．

表記類似度 $S'(w,o)$ に対して，文脈類似度 $C(w,o)$ を重みとして加重平均をとることで，候補語 w に対する「派生らしさ」のスコア $DScore(w)$ を式 (5) を用いて計算する．

$$DScore(w) = \frac{\sum_{o \in O_n}(S'(w,o) \times C(w,o))}{\sum_{o \in O_n} C(w,o)} \tag{5}$$

n は $DScore(w)$ の計算に $S'(w,o) \times C(w,o)$ の上位何件まで用いるかを表し，O_n は $S'(w,o) \times C(w,o)$ を降順に整列した上位 n 位に入る既存オノマトペ o の集合である．

2.6 「オノマトペらしさ」のスコア

奥村らは，表 1 に示す品詞ごとのオノマトペに後接しやすい文字列を伴う候補語の用例が多い場合，その候補語はオノマトペであると判定している．

そこで，式 (6) を用いて表 1 の後接文字列を伴う候補語 w の出現頻度を計算し，後接文字列の特徴量 $O(w)$ とする．

$$OScore(w) = \frac{\text{後接文字列を伴う } w \text{ の出現回数}}{w \text{ の出現回数}} \tag{6}$$

式 (7) を用いて名詞に後接する文字列を伴う w の出現頻度 $PN(w)$ を計算する．$PN(w)$ が 90%以上になった際は，w はオノマトペではないと判断し，後接文字列の特徴量 $O(w)$ を 0 とする．特徴量 $O(w)$ を「オノマトペらしさのスコア」$OScore(w)$ とする．

$$PN(w) = \frac{\text{名詞に後接する文字列を伴う } w \text{ の出現回数}}{w \text{ の出現回数}} \tag{7}$$

2.7 「派生オノマトペらしさ」のスコア

式 (8) を用いて，候補語 w に対する「派生らしさ」のスコア $DScore(w)$ と「オノマトペらしさ」のスコア $OScore(w)$ の積を取ることによって，w に対する「派生オノマトペらしさ」のスコア $Score(w)$ を計算する．

$$Score(w) = DScore(w) \times OScore(w) \qquad (8)$$

3 評価実験

3.1 方法

辞書 [4] に掲載されている 4,514 種類のオノマトペを既存オノマトペとして用い，Yahoo!知恵袋 API を用いて既存オノマトペを含む質問と回答を収集し，コーパスとして利用した．「AっB り」型の音韻パターンから 11,343 種類の候補語を生成し，既存オノマトペと同様に Yahoo!知恵袋 API を用いて候補語を含む質問と回答を収集した．

収集したコーパスから既存オノマトペおよび候補語を検索し，各々の既存オノマトペと候補語に対して用例文を取得した．用例文が 1 つも存在しなかった既存オノマトペおよび候補語を除くと，3,430 種類の既存オノマトペと 578 種類の候補語が残った．

各候補語に対して「派生オノマトペらしさ」のスコア $Score(w)$ を計算し，候補語をスコアの降順に整列した．上位 100 件についてコーパス中の用例を調べ，「オノマトペとして用いられている用例文が一つでも存在するか」という観点で正解判定を行った．さらに，$Score(w)$ を計算するために使用する特徴量 $S'(w,o)$，$C(w,o)$，$O(w)$ の組み合わせを変えて実験を行った．

奥村らの手法を比較対象とした．奥村らの手法は，候補語に表 1 に示した品詞ごとの後接文字列をつけてコーパス内で検索を行い，いずれかの品詞に対する検索結果が閾値を超えていれば，候補語をオノマトペであると判定する．本手法との比較のため，奥村らが検索結果が多い候補語をオノマトペらしい語と判定していることに基づき，奥村らの手法で抽出した候補語に対して順位付けを行った．まず，収集したコーパスに対して候補語に後接文字列をつけて検索を行い，品詞の種類ごとに検索結果を件数で降順に整列した後に，それぞれの上位 100 件をオノマトペである可能性が高い候補語として抽出した．次に，各品詞から抽出したオノマトペである可能性が高い候補語に対して品詞を考慮せず全品詞の後接文字列をつけてコーパス内で再度検索を行った．検索結果を件数で降順に整列した上位 100 件について本手法と同様の正解判定を行った．

3.2 結果

奥村らの手法に関する実験は 100 種類中，35 種類がオノマトペとして判定され，正解率は 0.35 であった．品詞の種類ごとに抽出したオノマトペである可能性が高い候補語に関しても，同様の正解判定を行った．品詞別の正解率は表 2 に示す．左の列は対象の品詞を表し，右の列は正解率を表す．

表 2: 奥村らの結果

品詞	正解率
サ変名詞	0.40
タル形容詞	0.40
ナ形容詞	0.28
ナノ形容詞	0.40
名詞	0.22
副詞	0.54

表 3: 本手法の結果

手法名	正解率		
	n=1	n=3	n=5
S	0.41	0.43	0.45
C	0.32	0.33	0.33
O	0.55		
SC	0.42	0.55	0.62
SO	0.58	0.53	0.54
CO	0.44	0.46	0.46
SCO	0.49	0.49	0.50

本手法に関する実験結果を表 3 に示す．一番左の列は，使用した特徴量による手法の名称を表す．二番目以降の列は，$DScore(w)$ の計算に $S'(w,o) \times C(w,o)$ の上位 n 件まで用いた際の正解率を表す．手法の名称に関して，S は特徴量 $S'(w,o)$，C は特徴量 $C(w,o)$，O は特徴量 $O(w)$ を $Score(w)$ の計算に用いたことを表す．

本手法で最も高い正解率は，$n = 5$ で SC を用いたときの 0.62 であった．

本手法でうまく抽出できた派生オノマトペとして，「かっぽり」「ねっちゃり」「ぺっちゃり」「もっそり」などが存在した．それぞれ「かっぽりと被るニット帽」「ねっちゃりしたクッキー」「髪がぺっちゃりしてる」「もっそり動きます」のような用例で用いられ，確かにオノマトペとして用いられていた．

本手法と奥村らの手法に関して比較を行った．C 以外の手法は全て奥村らよりも正解率が高い．O は奥村らと同様に後接文字列のみからスコア計算を行っているのに，奥村らよりも正解率が高い．これは，検索結果の頻度を用いたことで，用例数が少ない新しく使用し始められたばかりの派生オノマトペを抽出できたからである．

表記類似度の特徴量と文脈類似度の特徴量を組み合わせることが派生オノマトペの抽出に有効であるかという観点で比較を行った．SC と S，SC と C を比較すると，全ての n において正解率が向上した．これにより，「派生オノマトペと派生元のオノマトペは表記と文脈において類似する」という仮説の妥当性が分かる．

表記類似度の特徴量と後接文字列の特徴量を組み合わせることが派生オノマトペの抽出に有効であるかという観点で比較を行った．SO と S を比較すると，全ての n において正解率が向上した．一方で，SO と O を比較すると，$n = 1$ においては正解率が向上したものの，$n = 3$ と $n = 5$ においては正解率が低下した．

文脈類似度の特徴量と後接文字列の特徴量を組み合わせることが派生オノマトペの抽出に有効であるかという観点で比較を行った．CO と C を比較すると，全ての n において正解率が向上した．一方で，CO と O を比較すると，全ての n において正解率が低下した．

最後に,「派生らしさ」のスコアと「オノマトペらしさ」のスコアを組み合わせることが派生オノマトペの抽出に有効であるかという観点で比較を行った．SCO と SC を比較すると，$n = 1$ においては正解率が向上し，$n = 3$ と $n = 5$ においては正解率が低下した．SCO と O を比較すると，全ての n で正解率が低下した．

3.3 誤り分析

今回，最も正解率が高かった $n = 5$ とした際の SC に関して誤り分析を行った．この手法で上位 100 件以内にあり，かつオノマトペではないと判定された 38 件について誤りを分類したところ 4 種類に分けることができた．

一つ目の誤りは，動詞の全体または一部が含まれる語を抽出した点に起因し，25 件存在した.「つっきり」「しょったり」「ずったり」などが該当した．以下，それぞれについて考察する．

「つっきり」は,「商店街をつっきりながら」のように,「突っ切る」という動詞として用いられていた．

「しょったり」は,「酸素チューブを背中にしょったり」のように,「背負う＋したり」として用いられていた．

「ずったり」は,「内股でたまに引きずったり」のように,「引きずる＋したり」の一部として用いられていた．

「しょったり」と「ずったり」に対して $DScore(w)$ の計算に用いた上位 5 件の既存オノマトペを調べた．

「しょったり」に対して既存オノマトペ 1 位であった「しったり」は「知ったり」「走ったり」と用いられ，オノマトペとして用いられない場合が多かった．

「ずったり」に対して既存オノマトペ 1 位であった「すったり」は「擦ったり」や「吸ったり」と用いられ，オノマトペとして用いられない場合が多かった．

「しったり」と「すったり」はどちらもオノマトペとしての用例はほとんど存在しなかった．これにより動詞と動詞の間で文脈類似度を計算してしまい，既存オノマトペからの「派生らしさ」を計算できなかった．

二つ目の誤りは，名詞の全体または一部を抽出した点に起因し，6 件存在した.「はっとり」「まっこり」などが該当した．以下，それぞれについて考察する．

「はっとり」は $DScore(w)$ の計算に用いた既存オノマトペの 1 位が「はっ」であった.「はっとり」に「はっ」が含まれることで，用例文が重複し文脈類似度が高く計算された．

「まっこり」は「にっこり」「もっこり」などと共起しやすかった.「まっこり」に対して $DScore(w)$ の計算に用いた既存オノマトペを調べた．1 位と 2 位が「もっこり」と共起する「ぽっこり」「ぼっこり」であり，4 位には「にっこり」自体が入っていた．

残り 4 件の誤りは，用例が少なく文脈類似度が綺麗に計算できなかった．

三つ目の誤りは，オノマトペではない副詞を抽出した点に起因し，4 件存在した.「いっぱり」「かっなり」「だってり」「やっぱり」が該当した．

これらの語はそれぞれ「いっぱい」「かなり」「だったり」「やっぱり」という意味で用いられ，用例数はそれぞれ 22 文，28 文，20 文，20 文であった．用例が少なく文脈類似度が綺麗に計算できなかった．

抽出されなかった「AっB り」型の候補語に「だったり」「やっぱり」が存在した．これらの語は用例数がそれぞれ 1,008 文，1,004 文であった．用例が多いため,「だってり」「やっぱり」と同じ意味で用いられているにもかかわらず抽出されなかった．

最後に，上記以外のオノマトペではない語を抽出した誤りが 3 件存在した.「れっきり」「なっかり」「うっちゃり」が該当した．これらの語はそれぞれ「それっきり」「できなっかりして」「土俵際のうっちゃり」のように用いられていた．以下，それぞれについて考察する．

「れっきり」「なっかり」は用例が少なく文脈類似度が綺麗に計算できなかった．

「うっちゃり」は相撲の決まり手であり，$DScore(w)$ の計算に用いた既存オノマトペの 1 位である「ぽっちゃり」と文脈類似度が高く計算された．

4 おわりに

本研究は，派生オノマトペが造語される過程に着目することにより，新語の派生オノマトペを抽出する手法を提案した．より高い精度を目指すために,「オノマトペらしさ」のスコアや文脈類似度の計算方法を改良する必要がある．今回は「AっB り」型の音韻パターンを持つ派生オノマトペしか対象としていないため，様々な音韻パターンを持つ派生オノマトペも対象にする必要がある．

謝辞

本研究の一部は，科学研究費補助金基盤研究（B）（課題番号 22300050）によって実施された．

参考文献

[1] 奥村敦史, 齋藤豪, 奥村学. Web 上のテキストコーパスを利用したオノマトペ概念辞書の自動構築. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2003, No. 23, pp. 63–70, 2003.

[2] 中島正貴, 藤井敦. 造語の過程に基づく複合オノマトペの検出手法. 言語処理学会第 18 回年次大会発表論文集, pp. 69–72, 2012.

[3] 大野純子. 現代短歌・俳句に見る新語オノマトペ : 既存のオノマトペからの派生をとりあげて. 大正大學研究紀要. 人間學部・文學部, Vol. 94, pp. 1–13, 2009.

[4] 小野正弘（編）. 擬音語・擬態語 4500 日本語オノマトペ辞典. 小学館, 2007.