造語の過程に基づく複合オノマトペの検出手法

中島 正貴† 藤井 敦‡
† 東京工業大学工学部情報工学科
‡ 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

オノマトペとは擬音語と擬態語の総称である．不特定多数の人間がネットに情報発信するようになった昨今，新しい語が生まれる機会が多くなった．感情的な表現をするのに不可欠なオノマトペは，新語が創作される頻度が高い．オノマトペを正しく解析するためには，オノマトペの辞書を整備する必要がある．

奥村ら [2] は，Web コーパスを用いてオノマトペの概念辞書を自動構築する研究を行った．この研究では，オノマトペである可能性のある文字列 (以下，**候補語**と呼ぶ) を作成し，web から用例を収集したのち，候補語がオノマトペであるかを判定した上で辞書を構築する．このとき「音韻の規則性」を利用して候補語を作成しており，この規則性に添った語に限れば高い精度で取得できる．しかし規則性には限りがあり，網羅性は十分でない．

大野 [1] は新語のオノマトペを分類する中で**複合オノマトペ**という概念を提唱している．これは,「がぼがぼ」と「ごぼごぼ」を複合して「がぼごぼ」と言うように，複数の既存オノマトペから創作されるオノマトペである．

オノマトペの中で数が多いのは,「さらさら」,「つやつや」,「ちょろちょろ」など二つの音韻を繰り返したオノマトペである．本研究ではこの種のオノマトペを**反復オノマトペ**と呼び，2 つの反復オノマトペによる複合オノマトペを検出することを目的とする．

2 複合オノマトペの検出手法

2.1 概要

本研究で提案する手法は，反復オノマトペの一覧を入力として，複合オノマトペ候補語の順位付きリストを出力とする．順位付けに使うスコアを計算するためには,「オノマトペらしさ」の指標となる特徴量が必要である．そこで，複合オノマトペが造られる過程を考えることで特徴量とその計算方法を提案する．

「ゆるふわ」を例に複合オノマトペが造られる流れについて仮説を立てる．まず,「ゆるゆるで，ふわふわな」のように並列して用られるようになり，このうち助詞が抜けて結合すると「ゆるゆるふわふわ」という 1 つのオノマトペが形成され，さらに短縮されて「ゆるふわ」という形で用いられるようになると考えられる．この流れを 1 つのアルファベットが 1 音節を表すとして一般化すると，

$$ABAB + XYXY \rightarrow ABABXYXY \rightarrow ABXY$$

と表せる．このような場合，本稿では ABAB と XYXY を**親**と呼び，ABABXYXY を**中間状態**，ABXY を**子**と呼ぶ．また，ABXY と XYAB のように親の順序を入れ替えた関係を**兄弟**と呼ぶ．以上を図 1 にまとめる．

ただし，親の組み合わせによっては，オノマトペの可能性が低い候補語が生まれる．たとえば「ぐちょぐち (ぐちょぐちょ＋ぐちぐち)」は「ぐちょぐちょ」の一部としての出現が殆どであり,「ぐちょぐち」自体は非オノマトペである．そのため，AB と XY に文字列の包含関係がある語は，候補語を生成する段階で除外する．

以下，**2.2**〜**2.6** でオノマトペのスコア計算に用いる特徴量とスコア計算について説明する．

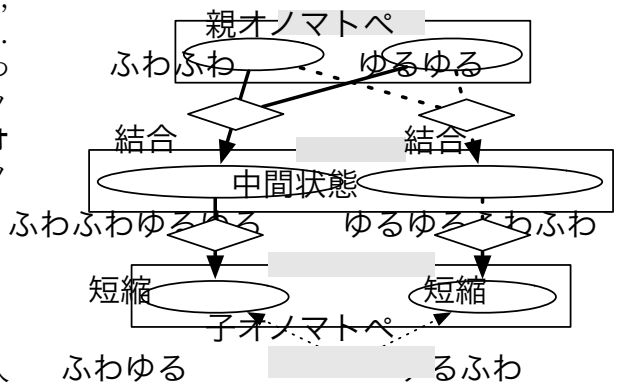

図 1: 複合オノマトペ造語の流れ

2.2 親と子の類似度

親と子は意味的に類似している可能性が高く，親と子は似たような文脈で用いられることが予想される．そこで，親と子の間で文脈類似度を計算して**特徴量 P(Parents)** とする．親と子を共起語の重みを要素とするベクトルで各語を表し，重みとして自己相互情報量を用いる．

まずコーパス全体に対し MeCab で形態素解析を行う．親オノマトペとその共起語の組み合わせ，および子オノマトペ候補語とその共起語の組み合わせに対し，それぞれ自己相互情報量 (PMI) を求める．ここで，ある語 W に対する共起語 C の $PMI(W,C)$ を式 (1) で計算する．

$$PMI(W,C) = \log \frac{p(W,C)}{p(W)p(C)} \tag{1}$$

$p(W,C)$ は W と C の共起確率であり，$p(W)$ と $p(C)$ はそれぞれ W と C の出現確率である．

二つの親および子がそれぞれベクトルで表されているとき，子と親のベクトルがなす角度の余弦を片親ずつ求め，相加平均をとった値を特徴量 P の値とする．

ただし，今回は P を計算する際に，子オノマトペ候補語としてカタカナの語のみを用いた．予備実験の段階で，新しいオノマトペはカタカナで用いられ始める傾向が強く，ひらがなで出現する場合は偶然の一致である場合が多かったからである．たとえば文 1 で「ぱりさく」は偶然の一致であるが，文 2 で「パリサク」はオノマトペとして用いられている．

文 1)　やっぱりさくらでしょうか？

文 2)　片栗粉をまぶしてパリサクに揚げる

2.3 中間状態

中間状態は子オノマトペに比べて全体的に出現頻度が高くない．そのため，中間状態の文脈類似度の計算は難しい．しかし一方で，その出現頻度自体に価値があると見ることもできる．そこで今研究では，中間状態の出現頻度を**特徴量 I(Intermediate State)** の計算に用いる．

ある中間状態のコーパスでの出現頻度が A であるとき，対応する候補語の特徴量 I を式 (2) に従って計算する．

$$I = \frac{2}{1 + e^{-mA}} - 1 \tag{2}$$

これはシグモイド関数を変形した式である．A を非負の実数である変数とするとき，値域が [0,1] となるよう調整した．m はグラフの傾きを制御するパラメタである．

2.4 字種による偏り

「すること (するする+ことこと)」や「よろしく (よろよろ+しくしく)」などのように，カタカナに対してひらがなでの出現頻度が顕著に高い語は偶然の一致である可能性が高い．そこで，式 (3) のように「ひらがなでの出現頻度」と「カタカナでの出現頻度」の比を計算し，**特徴量 C(Character Type)** とする．

$$C = \frac{2}{1 + e^{-n\frac{K}{H+1}}} - 1 \tag{3}$$

H と K はそれぞれひらがなとカタカナでの出現頻度であり，n は傾きを制御するパラメタである．H に 1 を足しているのは，分母が 0 にならないようにするためである．

2.5 候補語のスコア計算

候補語の順位付けに用いるスコアとして，特徴量 P, I, C の線形和を求める．P，I，C の値域は全て理論上 [0,1] であるものの，予備実験を行った結果，P は 1 より十分に低い値をとる傾向があった．そのため，I，C と P を等価に扱うためには値域を調整する必要がある．全ての候補語に対する P のうち，最大値が P_{max} であるとしたとき，P' を式 (4) のようにして求める．

$$P' = \frac{P}{P_{max}} \tag{4}$$

このとき P' の値域も [0,1] である．そして P'，I，C の線形和による候補語のスコア F を式 (5) のように求める．

$$F = \alpha\ P' + \beta\ I + \gamma\ C \tag{5}$$

α，β，γ でスコアの重みを調整することができるが，今回は 0 もしくは 1 とした．

また，兄弟語のスコアを自身の**特徴量 S(Sibling)** として用いるため，S の値を兄弟語の F の値とする．

3 評価実験

3.1 方法

今回は辞書 [3] に掲載されている「意味分類別索引」で出現する 505 種類の反復オノマトペを親として用いた．この 505 種類から 2 つの異なるオノマトペを選んで組み合わせ，中間状態および子オノマトペの候補語をともに 254,520 種類ずつ生成した．事前調査の結果，多くのオノマトペを含んでいた Yahoo!知恵袋をコーパスとして使用した．しかし手元にアーカイブがなかったため Yahoo!知恵袋 API を利用して，親，中間状態，子の候補語を含む質問と回答を収集した．

次に，収集した文章から各候補語に対して特徴量 P, I，C を計算し，α，β，γ の値の組み合わせを複数通り試行してスコア F を求めた．この F でソートしたときの上位 100 件についてコーパス中の用例文を調べ,「オノマトペとして用いられている用例文が一つでもあるか」という正解判定を行った．また，S によるソートを行って上位 100 件を取り出し，スコア F と同様に正解判定を行った．

奥村ら [2] も候補語がオノマトペかどうか判断する方法について提案している．この方法が複合オノマトペに応用可能かどうかについても検証を行う．奥村らの手法は，候補語にいくつかの後節文字をつけて検索を行い，

いずれかの結果が閾値を超えていればオノマトペであるというフラグを立てる．なお，フラグはオノマトペの品詞によって異なり，全部で 6 種類ある．これに従い，本実験で収集したコーパスに対して，候補語に後節文字をつけて検索を行った．そしてフラグの種類ごとに件数でソートし，それぞれの上位 100 件にオノマトペがいくつ含まれるか確認した．

3.2 結果

本実験を通して取得できたオノマトペには大きく分けて 3 種類あったので，それぞれについて説明をする．

1. **新オノマトペ**
   オノマトペ辞書 [3] には載っていない語．「つるてか」や「さらつや」などが該当する．
2. **既存のオノマトペ**
   辞書 [3] に掲載されている語．「めちゃくちゃ」や「びしばし」などが該当する．
3. **誤記の可能性があるオノマトペ**
   打ち間違いによる偶然なのか故意に使っているのか判断が難しい語．「ぽろぼろ」や「かさがさ」などが該当する．

まず，奥村らの手法で実験を行った．調査をした 600 件のうち，11 種類の既存オノマトペを検出できた．しかし，新オノマトペは 1 件も検出できなかった．

原因は 2 つあると考える．まず，奥村らの手法が本来対象としている候補語と今回の候補語では生成方法が異なるため，性質が異なる可能性があることである．そして，「新語」である以上，その出現頻度が低い傾向にあり，検索結果の件数によるソートでは新オノマトペが上位にくることが難しいということである．

次に，今回提案する手法で実験を行った．なお，シグモイド関数のパラメタの値には $m = 0.05$，$n = 10$ を用いた．表 1 にその結果を示す．左から，使用した特徴量による手法の名称，1〜50 位でのオノマトペ数，1〜100 位でのオノマトペ数を表す．記号◎，○，△はそれぞれ新オノマトペ，既存オノマトペ，誤記の可能性のあるオノマトペと対応する．

表 1: 評価実験の結果

	1-50				1-100			
	計	◎	○	△	**計**	◎	○	△
P	**32**	12	14	6	**48**	17	17	14
I	**37**	34	3	0	**60**	51	8	1
C	**0**	0	0	0	**0**	0	0	0
PC	**29**	10	13	6	**43**	15	15	13
IC	**41**	36	5	0	**66**	53	10	3
PI	**38**	29	9	0	**67**	49	14	4
PIC	**46**	34	12	0	**76**	52	14	10

まず手法 P では，◎，○，△のいずれもバランスよく検出していたため，P は複合オノマトペ全体に有効であることがわかった．◎の語は新語という性質から出現頻度が低いものも多いため，文脈類似度が低く計算されがちである．そのため，P は相対的に◎よりも○を検出するのに有効であると考えられる．

手法 I では多くの◎を検出したものの，○と△は殆ど検出できなかった．この結果から，「子オノマトペは中間状態を経て造られる」という仮説に一定の信憑性を与えることができる．

手法 C ではオノマトペを 1 件も検出することができなかったものの，手法 I と手法 IC，または手法 PI と手法 PIC を比較すると，C は他の特徴量と組み合わせることによって一定の効果があることが分かる．PI に C を加えることで 100 位圏内からの語の出入りは 24 件あり，そのうち出た語には 6 件の◎が，入った語には 9 件の◎と 6 件の△が含まれていた．これは，◎は出現頻度が低いため，出現頻度が僅かに変化するだけで C の値が大きく変動するからだと考えられる．特にひらがなでの出現頻度が 0 であるものは全て C=0 となるため，その影響で順位を下げていた．

最後に，特徴量 S を手法 PIC での兄弟語のスコアとして正解判定を行った．これを手法 S と呼ぶ．手法 S による上位 100 件のうち，手法 PIC で上位 100 件に現れた語と重複している語が 36 件存在しており，コーパス中に用例が 1 度も出現しない語も 21 件存在したので，これを除外した．残る 43 件について正解判定を行ったところ，◎が 9 件，○が 1 件，△が 6 件，計 16 件のオノマトペを検出できた．

3.3 誤り分析

手法 PIC で正解判定をする際に，上位 100 位以内にあり，かつオノマトペではないと判断した 24 件について誤りの原因を分類したところ，3 種類に分けることができた．

まず，I の値が高く，子オノマトペ自体の出現頻度が低い語である．つまり「中間状態としては使われるけれども，短縮されるまでには至っていない」ということであり，コーパスの範囲を広げれば既に子オノマトペとして使われている可能性があり，近い将来子オノマトペとして用いられる可能性もある．「ちくずき」「うずむら」「すらすい」「どきはら」「むかいら」「かりがり」「かすすか」「しくちく」「すかかす」「ぺろれろ」「ぬるねば」「くらふら」「にやにた」「ぴかきら」の 14 種類である．

次に，ひらがなに比べて，著しくカタカナでの出現頻度が高い語である．特徴量 C の式において H と K を入れ替えた値を使えば，順位が落ちることが期待できる．ただしその際は，カタカナで使われる場合が多いオノマトペもまた存在するということに留意しなければならない．また，これらの語は共通して P も高い値を示してい

表 2: 自己相互情報量の高い単語

ABXY	ABAB				XYXY			
	1	2	3	4	1	2	3	4
さらつや	しっとり	髪	髪の毛	あこがれる	髪	あこがれる	髪の毛	ブロー
がさごそ	くちびる	ポップコーン	物音	カバン	あさる	戸棚	カバン	物音
ぶらちら	ブルマー	ひも	路地	海岸	胸元	谷間	凝視	そそる
ぼてごろ	内野	ゴロ	サード	空振り	硬式	ゴロ	サード	フライ
のこぎり	クワガタ	冬眠	無期	切断	金槌	類推	遊泳	ふち
がらくた	渇れる	崩れ落ちる	ガレージ	積み上げる	ハイヒール	唯	すもう	費やす

た.「ぶらちら」「ぼてごろ」「どくくら」「ぷりくら」「つやぐら」「ちびでぶ」「にこから」の 7 種類である．結果的には,「略語であるもののオノマトペでない」語が多かった.

最後に，原因のわからない語があった.「のこぎり」「がらくた」「ちょこらん」の 3 種類が存在する．これらは中間状態では一度も出現しない．しかし，P が平均より偶然高く計算されたため現れてしまったと思われる.

この 3 種類の誤りのうち後に挙げた 2 種類について，非オノマトペであるにも関わらず特徴量 P の値が高かったため，それぞれ実際にどのような共起語が特徴量 P の値を押し上げたのかを調べた．調査対象としては 2 種類の誤りからそれぞれ 2 語ずつ,「ぶらちら」と「ぼてごろ」，および「のこぎり」と「がらくた」を用いた．親に対する自己相互情報量と，子に対する自己相互情報量を共起語ごとに求めて積を取り，その値の大きい方から片親ごとに 4 件ずつ示したものが表 2 である．特徴量 P が高い値であったオノマトペ「さらつや」と「がさごそ」についても比較対象として載せている.

オノマトペである「さらつや」と「がさごそ」はともに両親とも同じような語と共起していることが分かる．一方で，非オノマトペである「のこぎり」,「がらくた」,「ぶらちら」は両親に共通する共起語が見当たらなかった．このことから,「候補語がオノマトペであるならば，どちらの親とも同じような語が共起する」という仮説が立てられる．この仮説の検証は今後の課題とする．一方で,「ぼてごろ」は「どちらの親とも同じような語と共起するものの，オノマトペではない」候補語である．これは「ぼてごろ」が「ボテボテのゴロ」という語の略語であるからだと考えている.

また，手法 PIC では上位 100 件中に 76 件のオノマトペが存在し正解率 76%であったのに対し，手法 S では，オノマトペが 16 件，非オノマトペが 27 件で 37%の正解率となってしまった．この原因は 3 つある.

まず，非オノマトペの兄弟語を参照しても非オノマトペであったということである．手法 S による非オノマトペ 27 件のうち，15 件が手法 PIC で非オノマトペとされていた語の兄弟語であった．逆に，手法 PIC で非オノマトペとされていた語の兄弟語にそもそもオノマトペは存在しなかった.

次に，新オノマトペに対して「既存のオノマトペは兄弟オノマトペを造りにくい」傾向があるとことである．例えば「ぺちゃくちゃ」に対して「くちゃぺちゃ」とはあまり言わないということである.

最後に，手法 S で検出した語は出現頻度が低い傾向があったことがある．手法 S による非オノマトペ 27 件のうち，出現頻度が 10 件未満であった語は 14 件あった．コーパスの規模を大きくすれば複合オノマトペとしての用法が得られる可能性がある.

4 おわりに

本研究では，複合オノマトペとその成り立ちに着目することで，オノマトペの新語を検出する手法を提案した．より高い精度を目指すため，親同士の関係性など他の特徴を用いる必要がある．また，今回は複合オノマトペの中でも反復型同士によるものしか扱っていないため，今後は他の組み合わせによる複合オノマトペへ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] 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, , 2009-03.

[2] 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, 2003-03-06.

[3] 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.