国際音声記号を素性とした 3文字以下の未知のオノマトペ自動抽出手法の提案

池田祐一[†1]　阪本浩太郎[†2]　渋木英潔[†3]　森辰則[†3]

†1 横浜国立大学 理工学部　†2 横浜国立大学 大学院 環境情報学府
†3 横浜国立大学 大学院 環境情報研究院
E-mail: {ikd8e3a1,sakamoto,shib,mori}@forest.eis.ynu.ac.jp

1　はじめに

日本語におけるコミュニケーションの中で擬音語，擬態語といったオノマトペは，表現を豊かにするためのものとして様々な場面で用いられている．例えば，患者が医者に痛みの質や程度を伝える場合 [1, 2] や，新しく作成したデザインの印象 [3, 4] や素材の触感 [5] などを伝える場合などである．他にも，第二言語としての日本語学習者に対するオノマトペ学習支援 [6] や，オノマトペを利用したレビュー文の評価極性推定 [7] などオノマトペに関する研究は多い．我々は，地方議会会議録を対象とした政治情報システムの構築 [8] に取り組んでいるが，地方議会会議録においてもオノマトペが豊富に使用されており,「きちっと進めていきたい」のように政策の推進や「はっきりと述べてください」のように厳格な言及を求めるために用いられていることが指摘されている [9]．こういったオノマトペを正しく抽出することができれば，発言議員の政策に対する積極性などを判断するのに役立つと考えられる．

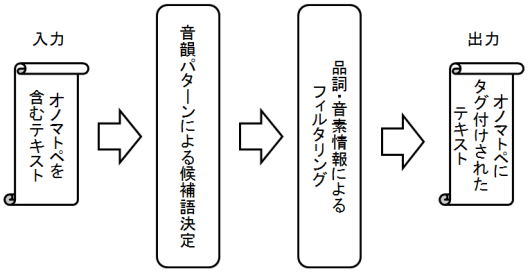

図 1: 処理の流れ

本研究では，図 1 のように，まず入力テキスト中の文字列からオノマトペの候補となる文字列（オノマトペ候補）を決定する．オノマトペ候補を認識する手法として，オノマトペを辞書に登録してそれを利用する方法もあるが，オノマトペには新語や造語が多く存在するため，すべてのオノマトペをあらかじめ辞書に登録することは不可能である．そこで本手法では音韻パターンに基づいてオノマトペ候補を決定する．次に認識されたオノマトペ候補とその前後 n 文字における各文字の音素情報と各形態素の品詞を手がかりとして，オノマトペ候補がオノマトペかどうかを判断する．

本手法で用いる音素は国際音声記号に基づくものであり，この点が本手法の特徴の一つである．

また，比較的解析が難しい 3 文字以下オノマトペのうち，未知のものに対応することを目的としている点も従来研究には見受けられない．

2　従来研究

JUMAN[1] では，笹野ら [11] や勝木ら [12] の研究成果を利用して，パターン一致によりオノマトペの候補語を決定する．ここで用いられるパターンはいずれも 4 文字以上のものである．

木村ら [13] は，地方議会会議録コーパスにおけるテキスト中からオノマトペを適切に抽出する手法として，形態素解析結果に基づく規則と，構文解析結果に基づく規則を用い，オノマトペを 4 文字以上と 3 文字以下の 2 種類に分けそれぞれに異なる規則を適用した．3 文字以下のオノマトペの抽出に効果を発揮する手法であり，3 文字以下のオノマトペの場合目的のオノマトペを内部辞書にあらかじめ登録するだけの手法と比較して適合率が 39.5 %から 85.4 %と大きく向上した．対象とするオノマトペがあらかじめ決まっている点が本手法と異なる．

3　田守・スコウラップによる分類

提案手法では，オノマトペの持つ基本的なパターンに合致するものをオノマトペ候補とすることで解析を行う．

田守･スコウラップによる，オノマトペの音韻パターン分類 [14] を利用してオノマトペ候補を認識する．子音「C」，母音「V」，撥音「N」，促音「Q」，CV のペアに付随して現れる「り」を表す「ri」といった表記を用いて表される田守・スコウラップの分類には「どん」のような CVN 型,「ばたり」のような CVCVri 型など 16 種類がある．本稿の実験では 3 文字以下のオノマトペを対象とした研究の手始めとして,「ばたっ」や「ぴちゃっ」のような CVCVQ 型のみを対象に実験を行った．

音韻パターンをテキスト上の表層パターンと対応させるため，CV の組を小文字を除くひらがなカタカナ 1 文字，または大文字と「っ」を除く小文字の組み合わせに対応させ，C を伴わない V を「ー」，N を「ん」または「ン」，Q を「っ」または「ッ」とそれぞれ対応させた．

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

4　品詞および音素情報の利用

3 節の分類を利用して抽出したオノマトペ候補語がオノマトペか否かを判定するために，素性として品詞や音素情報を利用する．オノマトペの周辺，特にオノマトペの直後に付随して現れる語の品詞を情報として利用する手法は内田ら [15] の研究などで見受けられるが，音素情報を利用した情報は従来手法にないものである．オノマトペ候補語の判定に品詞および音素情報を用いる理由は以下のとおりである．

CVCVQ 型は直後に助詞を伴うことが多いなど，オノマトペはパターンごとに直前直後の共起しやすい品詞が存在する．また，例えばオノマトペ「がくっ」を含む文「なにかがくって落ちた」という例文においてオノマトペ「がくっ」が助詞「が」＋動詞「食って」の一部，として解析されるように,「がくっ」は助詞＋動詞という誤解析になりやすい，といったオノマトペの解析され方の特徴も考えられる．これらを考慮に入れた解析を行うために，オノマトペ候補とその前後の品詞列情報を利用する．品詞の付与には形態素解析器 MeCab を利用した．

また，オノマトペの新語や造語に対して受け取り手がそれをオノマトペと判断するための要因の一つとして音素による情報が重要である [10] という事実から，音素による情報も利用する．音素情報として国際音声記号 (IPA)[2] を用いて表 1 の分類情報を利用する．

表 1: IPA による分類

母音, 破裂音, 破擦音, はじき,
鼻音, 摩擦音, 接近音

これらのうち各ひらがなカタカナ 1 文字が持つ子音の調音方法 (あ行の文字は母音) を付与する．

5　機械学習手法による解析ルール獲得

本研究では，様々な素性を利用した解析を行うため，それらを用いた解析のためのルール構築には機械学習を用いる．本稿では，チャンキングツール Yamcha[3] による SVM を用いた系列ラベリング手法によって機械学習を行った．

5.1　データフォーマット

Yamcha の入出力および機械学習に用いる訓練データのフォーマットは以下のようになる

上の例は本稿の実験で用いた訓練データの一部である．本実験では文字によるチャンクを利用し，タグはオノマトペの出現位置を表す IOB2 フォーマットのタグである．品詞は該当する形態素中の文字すべてに付与し，音素情報は文字ごとに付与する．先にも述べたとおり，3 文字以下のオノマトペに対する手法の有効性の予備的な実験として，今回の実験ではパターンは CVCVQ のみを利用する．パターンに合致した箇所に含まれる文字すべてに合致するパターン名を素性として付与し，今回の実験ではそれ以外のパターンに該当する箇所は無視した．

表 2: Yamcha のデータフォーマット

すぐ	副詞	摩擦音	NONE	O
それ	副詞	破裂音	NONE	O
が	名詞	摩擦音	NONE	O
ばく	名詞	はじき音	NONE	O
っと	助詞	破裂音	NONE	O
折れ	名詞	破裂音	CVCVQ	BOnomatopoeia
てし	名詞	破裂音	CVCVQ	IOnomatopoeia
まっ	助詞	破裂音	CVCVQ	IOnomatopoeia
た	助詞	破裂音	NONE	O
。	動詞	NONE	NONE	O
。	動詞	はじき音	NONE	O
。	助詞	破裂音	NONE	O
。	動詞	摩擦音	NONE	O

5.2　素性の抽出と正解ラベルの付与

訓練データの作成は，[9] において作成された地方議会会議録コーパス中で形態素解析によってオノマトペとして解析された箇所に対して人手でオノマトペか否かを判定した結果をまとめた評価データ群をコーパスとして利用した．この評価データは，木村ら [13] の研究において利用されたものと同様である．形態素解析によってオノマトペと解析された箇所を 1 つ以上含む各々の文について，人手での判断が情報として付与されている．1 つ 1 つの評価データはオノマトペ 177 語ごとに分かれており，1 つの評価データ中には対象となるオノマトペに関する判定がなされた文が集められている．つまり,「ばくっ」に関する評価データ中には，形態素解析の結果オノマトペ「ばくっ」であると解析された箇所を含む文とそれに関する人手による判定の結果が集められており，そのような評価データが 177 語それぞれに 1 つずつ存在するということである．ただし，形態素解析によってオノマトペと解析された箇所の出現位置についての情報は付与されていない．

正解データの作成は利用するコーパスの情報から自動で行った．1 つの文中に形態素解析によってオノマトペとして解析された箇所が複数存在する場合それらのうちいずれか 1 つの判定結果のみ情報として保持するためこのような場合は複数箇所すべてに対してコーパスが保持する判定結果を適用させた．そのため実際の判定と異なる場合がある．今回の実験ではこれを誤差と捕らえたが，今後この誤りによる影響がどの程度のものであるかに関する判定と判定結果によるこの誤りの扱い方について検討する必要がある．

また，この方法では対象でないオノマトペにはタグを付与することができないため，そのことによる誤まりが含まれたデータとなっている．この誤りに関しても今後の検討が必要である．

コーパス中で学習に利用されるべき部分はオノマトペの付近のみであると仮定し，そこに含まれない部分における学習結果が解析に悪影響を与えないようにするため，オノマトペタグが付与された各部分の前後 5 文字を含めた 13 文字を 1 つの文として学習に用いた．つまり，評価データ中に「… ほんとうにばさっとして

[2] https://www.internationalphoneticassociation.org
[3] http://chasen.org/ taku/software/yamcha/

がくっときたので … 」という部分が含まれていた場合，オノマトペ「ばさっ」と「がくっ」に対して「ほんとうにばさっとしてがく」と「さっとしてがくっときたので」という二つの文が生成され順に並べられた構造の学習データを作るということである．

また，Yamcha の設定として前後 5 チャンク (本手法では文字チャンクなので前後 5 文字) までを利用し，文頭からラベリングを行う方法で学習を行った．

6 評価実験

6.1 実験方法

5.2 節の評価データの中で判定が行なわれている全 177 語のうち，CVCVQ 型は 25 個存在する．CVCVQ 型のオノマトペに関する評価データ 25 個を 24 : 1 に分け，それらから作成した訓練データ 24 個と入力データ 1 個のデータセットを，25 語それぞれの評価データが入力データとなるような組み合わせ 25 通り作成し交差検定を行った．コーパス中でオノマトペと判定されている箇所，オノマトペではない (非オノマトペである) と判定されている箇所それぞれについて，本手法によるタグ付けがオノマトペであると判定した箇所の個数，非オノマトペであると判定した箇所の個数を人手で確認した．

6.2 結果と考察

交差検定の結果は表 3 のようになった．

これをもとに計算されるシステム全体の再現率は 563/1520 = 37.0 %，適合率は 563/777 = 72.4 % である．3 文字以下のオノマトペを抽出するための他の手法と異なり，未知のオノマトペの抽出を可能にすることが本手法の目的であるため，再現率が重要となる．再現率に関しては比較対照がないため今回の結果をベースラインとし，今後の手法改善に取り組む必要がある．一方適合率に関して，今回利用した評価データ作成時に利用された形態素解析による適合率が 39.5 %，木村ら [13] の手法による適合率が 85.4 %であることと比較すると，本手法と形態素解析は対象とする語が複数ある (本手法は未知のオノマトペを含めた無数の語，形態素解析は辞書に含まれている語) もの同士の比較ができ，適合率の向上が行えている．それに対し対象とする語が 1 つである木村らの手法との比較では，再現率は低下している．

解析の質に悪影響を与えた原因について考察する．素性に用いた品詞による原因と考えられるものは 2 通りある．1 つは MeCab の内部辞書においてオノマトペ＋助詞「と」による副詞節が副詞 1 語として登録されている場合である．表中で「*」が 1 つ付随するオノマトペがこれに当たる．この場合オノマトペと後続する助詞「と」に副詞タグが付与される．もう 1 つが形態素解析の結果，オノマトペの後ろ 2 文字 (ここで「ちゅ」や「ふぁ」などは 1 文字と数える) ＋助詞「と」が副詞として解析されやすかったものである．表中で「*」が 2 つ付随するオノマトペがこれに当たる．この場合オノマトペの末尾 2 文字と後続する助詞とに副詞タグが付与される．いずれもオノマトペに付随した助詞「と」が助詞ではなく副詞のタグを付与されやすく，このようなケースでは比較的再現率が小さくなる傾向にあることが結果からわかる．よって CVCVQ 型のように助詞「と」を後ろに伴いやすい 3 文字以下のパターンに該当するオノマトペを解析する場合，パターンによって取り出した箇所の直後に「と」が来てかつそれが形態素解析器によって副詞の一部として解析されていたならば品詞を副詞に変更した状態で品詞付与を行うなどの工夫が有効であると考えられる．

また，CVCVQ に合致しており，該当箇所とその前後の品詞が同一であるが，IOB2 タグが付与されているものとされていないものがそれぞれ一定数以上存在するようなケースも存在した．これらは該当箇所前後の音素情報の違いによる影響が大きいと考えられる．音素情報は CVCVQ に合致した部分のみ有効であると考えられるため，前後の音素情報は含めない学習手法を検討する必要がある．

さらに，系列ラベリング手法では，CVCVQ に合致した部分以外も学習の対象となるが，本手法では音韻パターンによってあらかじめオノマトペの候補部分を決定するため，候補部分のみを学習の対象とする手法に変更するのが有効であると考えられる．

以上の考察から今後の課題として YamCha による系列ラベリング手法ではなく SVM を用いた学習方法を行う，音韻パターンに合致する部分のみを学習の対象とし，その前後の部分からは品詞情報のみを利用する，その直後に「と」を伴う場合は先に述べたような処理を施すといった改良を行っていく必要がある．

7 まとめ

今回は音韻パターンのうち CVCVQ 型のみを対象に実験を行い，3 文字のオノマトペに対するシステムの有効性を評価した．6.2 節で述べたような問題に関する検討を行うこと，その他の音韻パターンにおいても同様の評価実験を行うこと，また複数の音韻パターンを同時に利用して実験を行った際のシステムの有効性の評価を行うことが今後の課題である．さらに，音素情報がシステムに与える影響についてこれを利用す場合としない場合の比較により評価することが必要である．

参考文献

[1] 坂本真樹, 小野正理, 清水祐一郎. 痛みを表すオノマトペを用いた問診支援システム, 第 26 回人工知能学会全国大会口頭発表, 2N1-OS-8c-2, (2012).

[2] 上田祐也, 清水祐一郎, 坂口明, 坂本真樹. オノマトペで表される痛みの可視化, 日本バーチャルリアリティ学会論文誌, vol.18, no.4, pp.455–463, (2013).

[3] 飯場咲紀, 志賀彩乃, 坂本真樹. オノマトペによる色彩提案システム，第 26 回人工知能学会全国大会口頭発表，1M2-OS-8b-4, (2012).

[4] 土斐崎龍一, 飯場咲紀, 及川歩唯, 清水祐一郎, 坂本真樹. オノマトペによる画像色彩推薦，日本バーチャルリアリティ学会論文誌, vol.18, no.3, pp.357–360, (2013)

表 3: 交差検定による評価実験結果

コーパス	オノマトペ		非オノマトペ	
本手法	オノマトペ	非オノマトペ	オノマトペ	非オノマトペ
ばくっ	73	0	0	0
ばりっ*	1	0	47	12
ばさっ**	34	58	0	0
ぼやっ*	18	27	0	33
どきっ**	32	10	0	0
ふわっ*	0	36	0	0
がばっ	39	2	71	8
がくっ	49	2	0	2
がしっ	1	0	0	38
がたっ*	34	1	0	5
かちっ**	3	78	0	3
からっ*	6	4	9	18
きちっ*	0	102	0	1
ころっ	8	127	0	5
ことっ	0	0	87	39
もやっ	0	7	0	104
にやっ**	0	9	0	178
ぴちっ*	62	5	0	0
ぴしゃっ*	14	100	0	0
ぴたっ*	95	4	0	0
さらっ	69	94	0	1
しとっ	0	0	0	90
すかっ*	18	45	0	40
ずばっ	4	34	0	0
ずらっ	3	212	0	0

[5] 渡邊淳司, 加納有梨紗, 坂本真樹. オノマトペ分布図を利用した触素材感性評価傾向の可視化, 日本感性工学会論文誌, vol.13, no.2, pp.353–359, (2014).

[6] 楊碩, 橋本敬, 李冠宏, 李暁燕. 創作タスクによる日本語オノマトペのニュアンス学習方法に関する研究, 言語処理学会第 20 回年次大会発表論文集, pp.117–120, (2014).

[7] 五十嵐沢馬, 笹野遼平, 高村大也, 奥村学. オノマトペの音象徴を利用した評判分析, 言語処理学会第 18 回年次大会発表論文集, pp.715–718, (2012).

[8] 筒井貴士, 我満拓弥, 大城卓, 菅原晃平, 永井隆広, 渋木英潔, 木村泰知, 森辰則. 地方議会会議録コーパスの構築および政治情報システム構築を目標としたアノテーションの一提案. 自然言語処理, vol. 21, no. 2, pp. 125–156, (2014).

[9] 高丸圭一, 内田ゆず, 乙武北斗, 木村泰知. 地方議会会議録コーパスにおけるオノマトペ -出現傾向と語義の分析-, 人工知能学会論文誌, vol.30, no.1, pp.306–318, (2015).

[10] 戸本裕太郎, 中村剛士, 加納政芳, 小松孝徳. 音素特徴に基づくオノマトペの可視化, 日本感性工学会論文誌, vol.11, no.4, pp.545–552, (2012).

[11] 笹野遼平, 黒橋禎夫. 形態素解析における連濁および反復形オノマトペの自動認識, 言語処理学会第 13 回年次大会発表論文集, pp.819–822, (2007)

[12] 勝木健太, 笹野遼平, 河原大輔, 黒橋禎夫. Web 上の多彩な言語表現バリエーションに対応した頑健な形態素解析, 言語処理学会第 17 回年次大会発表論文集, pp.1003–1006, (2011).

[13] 木村泰知, 渋木英潔, 内田ゆず, 乙武北斗, 高丸圭一, 森辰則. 地方議会会議録におけるオノマトペの自動抽出手法の提案, 第 30 回ファジィシステムシンポジウム, (2014).

[14] 田守育啓, ローレンス・スコウラップ, オノマトペ—形態素と意味—, くろしお出版, (1999).

[15] 内田ゆず, 荒木健治, 米山淳, ブログ記事から抽出したオノマトペの多義性について, 第 24 回ファジィシステムシンポジウム, (2012).