音素単位の分散表現に基づくオノマトペ辞書構築法の提案

馬場　睦也† 　楠　 和馬† 　波多野賢治††

† 同志社大学大学院文化情報学研究科　〒 610–0394 京都府京田辺市多々羅都谷 1–3
†† 同志社大学文化情報学部　〒 610–0394 京都府京田辺市多々羅都谷 1–3
E-mail: †{baba,kusu}@mil.doshisha.ac.jp, ††khatano@mail.doshisha.ac.jp

あらまし　EC サイトや SNS 上のオノマトペを含む文章は，商品を実際に使用した者によって投稿された可能性が高く，評判分析においてオノマトペの検出は重要である．既存手法はオノマトペにおける音素構造の規則性に基づいた辞書構築法を提案しているが，オノマトペではない文字列も辞書に登録してしまう．この問題点を解決するためには，言語学者によって報告されている音素構造の規則性に対する制約が有効である．ここでいう制約とは，オノマトペを構成する音素の組合せを限定するものであり，音象徴に基づいている．音象徴とは，音素が表現する意味である．そこで，本研究では音素構造の規則性だけでなくオノマトペの特徴の一つである音象徴も考慮したオノマトペ判別モデルを構築する．判別モデルを評価するため，まずパラメタチューニングを行い，判別精度の良いモデルを構築する．最後に，提案手法の結果と既存手法の結果を比較し評価を行う．

キーワード　オノマトペ，音象徴，分散表現

1 はじめに

オノマトペとは，猫の鳴き声を表す「にゃー」のような語を意味する擬音語と，驚いた状態を表す「びくっ」のような語を意味する擬態語である [1]．特に日本語の場合，本のような商品の感想・評価が記述されているイーコマース（EC）サイト上のオノマトペを含む文は，実際に使用したユーザによって記述されている確率が高く，評判分析において有用であることが報告されている [2]．これより，デジタルマーケティングのようなソーシャルネットワークキングサービス（SNS）や EC サイト上の文から情報抽出を行う際，文にオノマトペが含まれているかどうかという点は重要であることが分かる．

英語のような文内で語ごとに空白で区切られている言語であれば計算機であっても調べることが可能である．しかし，日本語は文内に区切り文字がないため，形態素解析を行うことにより複数の単語に分割してからオノマトペが含まれているかを調べる必要がある．形態素解析は，文から形態素への分割および各形態素の品詞や活用形などを判定する処理であり，事前に形態素と呼ばれる意味を持つ最小単位である語がまとめられた辞書に基づいて文の分割を行う．以上の理由から，オノマトペである語が文に含まれているかどうかを調べるためには，オノマトペを集めた辞書が必要となる．

このため，オノマトペの辞書構築に関する研究がこれまでに行われている．その中でも最も有効だとされている手法として，言語学の一分野である音韻論で定義されているオノマトペの構造的規則（オノマトペパタン）を用いたオノマトペ辞書の構築方法がある [3, 4]．Okumura らの研究 [3] では，オノマトペパタンの一部を使用し，オノマトペパタンと一致する文字列を Web 上で検索した際に得られる文を用いたオノマトペ辞書の自動構築法が提案されている．しかし，この方法は Web 上の文に依存しているため，検索して得られる文が十分な数でなければ対応できない．また他の方法として Sasano らの手法 [4] は，筧・田守によって定義されているオノマトペパタン [5] の一部を使用し，形態素解析時にそのオノマトペパタンに一致する文字列のコスト調整を行う手法の提案している．しかし，この手法は Web 上のテキストでは良い判別精度が確認されているが，ドメインが異なれば文の表現や使用される語も異なるため，レビューや SNS 上の文に対しても同様に活用できるとは限らない．この理由は，オノマトペパタンに合致する文字列をそのまま辞書に追加することによって，オノマトペではない文字列までオノマトペとして辞書に登録されてしまうためである．

オノマトペには，オノマトペパタンをより詳細にし，オノマトペである語にみられる規則を定義する制約がある．つまり，制約はオノマトペパタン内のオノマトペではない文字列を取り除くことが可能になると考えられる．このため，制約はオノマトペであるかどうかの判別に有効である．またこの制約は，子音や母音などの音素が持つ意味である音象徴に基づいて言語学者によって定義されている．

そこで本研究では，音象徴に基づくオノマトペ辞書の構築法を提案する．音象徴を考慮することでオノマトペパタンやオノマトペパタンの表現する文字列を制限する制約を考慮した判別モデルの構築が可能になり，オノマトペではない文字列の辞書への混入を防ぐことが可能になると考えられる．これにより本研究では，オノマトペパタンの網羅性を維持しつつ，オノマトペではない文字列の辞書への追加を防ぐことを目的とする．また提案手法の有用性評価を行うため，まずパラメタチューニングを行い構築したオノマトペ判別モデルの当てはまりの良さを確認する．次に，最も判別精度が良かったオノマトペ判別モデルを用いた手法と Sasano らの手法 [4] の精度を比較し，提案手法の辞書構築法への有効性を評価する．

2 関連研究

本節では，まず 2.1 節にてオノマトペ辞書を構築する際に，基礎となっているオノマトペパタンについての説明を行う．次に，2.2 節および 2.3 節にてオノマトペパタンを用いたオノマトペ辞書構築法の概要とその問題点について説明を行う．そして，オノマトペパタンに課されている制約および制約の定義行う際，基礎となった音象徴について 2.4 節で述べる．

2.1 オノマトペパタン

オノマトペは，音韻論の観点から見た際，規則性がみられる [1]．音韻論とは，言語学の一分野であり，言語で用いられている音に対する研究として，単語や句，音節といった言葉を扱う上で使用される単位の観点から音との関係性を扱っている [6]．音韻論で扱う音の最小単位は，話者が発音しようとしている音を意味する音素であり，具体的には子音や母音，撥音，促音などである．例えば,「watashi（私）」を音素ごとに分割すると以下のようになる．

“ w ”，“ a ”，“ t ”，“ a ”，“ sh ”，“ i ”

音素単位で確認した際にみられる具体的なオノマトペの規則性として，主なものにモーラの観点からみた規則性がある．モーラとは，拍とも呼ばれ，音の長さを表す単位である．例えば,「わたし」をモーラ単位で分割すると次のように分けられ，3 モーラであるとされる．

“ わ ”，“ た ”，“ し ”

オノマトペのほとんどは，1 モーラである CV パタンもしくは 2 モーラである $CVCV$ パタンを基本形とするパタンに従うことが分かっている．C および V はそれぞれ子音および母音を意味しており，CV と $CVCV$ は子音と母音の組合せで表現できる文字列を意味する．これら二つの基本形は，このままオノマトペとして用いられることは少なく，促音である Q や撥音である N，繰返し表現などを用いたものが多い．具体的には，CVQ で表される「saQ（さっ）」，$CVCVN$ で表される「bataN（ばたん）」，$CVCV$ の繰り返しで表される「dotabata（どたばた）」などが挙げられる．現在，言語学者によって報告されているオノマトペパタンは 20 種であり [1,7,8]，その詳細は付録の表 A·1,A·2 に示す通りである．

2.2 Web 上のテキストを用いた手法

Okumura らは，言語学の音韻論で定義されているオノマトペパタンと Web 上のテキストデータを用いたオノマトペ辞書の自動構築法を提案している [3]．具体的には，まず，オノマトペパタンの一部を用いてオノマトペ候補語を生成する．次に検索エンジンを用いてオノマトペ候補語を調べ，検索該当件数が閾値に届いていない場合には，候補語から外し，閾値を超えていれば候補語として残す処理を行う．そして，候補語が含まれているテキストから，候補語の文脈情報を用いて品詞の推定を行う．ここでいう文脈情報とは，候補語の係り受け関係や品詞を意味する．この際，オノマトペがとるとされている品詞である可能性が閾値を超えなければ，その候補語は，外される．この手法は，オノマトペかどうかの判別を 83.6%で行うことが可能である．

しかし，この手法は，Web 上のテキストに大きく依存している点が問題であると考える．つまり，とあるオノマトペである語が含まれている文が Web 上で十分な件数確認されなければ，該当する語がオノマトペであってもオノマトペであると判別できない．

2.3 形態素コストの調整を行う手法

オノマトペパタンの一部を用いたオノマトペ辞書構築法は他にも提案されており，Sasano らの手法 [4] では，Web 上の文に対する検索結果に依存せず，オノマトペパタンの一部を用いて，それらと一致する文字列全てを辞書に含め，正しく形態素解析が行われるよう，形態素解析時使用される形態素コストを調整する方法がとられている．

しかし，Sasano らの手法 [4] で構築した辞書は，Web 上の文に対して精度が確認されているが，レビューや SNS 上の文には対応できない可能性がある．つまり，オノマトペではない語をオノマトペとして検出する可能性がある．この原因は，オノマトペではない文字列が辞書に含まれているためである．この問題を解決するためには，辞書からオノマトペではない文字列を取り除く必要がある．

2.4 音象徴に基づく制約

辞書からオノマトペではない語を取り除くためにはオノマトペパタンに一致する文字列がオノマトペであるかどうか判別しなければならないが，この判別基準として言語学者により定義されている制約がある．ここでいう制約とは，オノマトペパタンに課されている音素の共起および位置を限定するものであり，言語学者が実在するオノマトペを考察することで発見されている [8]．例えば，$CVCV$ パタンの C にはタ行を表す子音の “ t ” とダ行を表す子音の “ d ” が共起することはほとんどないとされている．言語学者は，このような制約を定義する際，音素の持つ意味である音象徴 [9] に基づいており，各音素の意味については，付録の A·3,A·4,A·6,A·5 にまとめた．先例をあげて説明すると，$CVCV$ パタンで “ t ”，“ d ” が共起しにくい理由は，これらの音素の意味が類似しているためである．

音象徴は，実在するオノマトペの音素列を基に言語学者によって定義されており，各音素に意味がある．具体的には，子音の一つである “ s ” は，“ suQ ”(すっ) や “ saQ ”(さっ) などに基づいて，抵抗のない表面や流動体などの意味があるとされており，母音の一つである “ a ” は，“ paQ ”（ぱっ）や “ paN ”（ぱん）などに基づいて，広がりの意味があるとされている．以上の理由から音象徴を考慮したうえでみられる規則性を基にオノマトペ辞書を構築することは，既存研究の問題点を解決する方法として有効であると考えられる．

表 1　Sasano らの手法 [4] が用いたオノマトペパタン

オノマトペパタン	具体例
S_1QS_2ri	mo$Q$sari，pe$Q$tyari
S_1S_2Qto	kariQto，sowaQto
2〜4 文字の繰り返し	gujogujo，uhauha

表 2　S_1 と S_2 の詳細

S_1	a, i ,u ,e, o, ka, ki, ku, ke, ko, sa, si, su, se, so ta, ti, tu, te, to, na, ni, nu, ne, no, ha, hi, hu, he, ho ma, mi, mu, me, mo, ya, yu, yo, ra, ri, ru, re, ro wa, ga, gi, gu, ge, go, za, zi, zu, ze, zo da, di, du, de, do, ba, bi, bu, be, bo, pa, pi, pu, pe, po
S_2	a, i, u, e, o, ka, ki, ku, ke, ko, sa, si, su, se, so ta, ti, tu, te, to, na, ni, nu, ne, no, ha, hi, hu, he, ho ma, mi, mu, me, mo, ya, yu, yo, ra, ri, ru, re, ro wa, ga, gi, gu, ge, go, za, zi, zu, ze, zo da, di, du, de, do, ba, bi, bu, be, bo, pa, pi, pu, pe, po kya, kyu, kyo, sha, shu, sho, tya, tyu, tyo, nya, nyu, nyo hya, hyu, hyo.mya, myu, myo, rya, ryu, ryo, gya, gyu, gyo ja, ju, jo, bya, byu, byo, pya, pyu, pyo

3　提 案 手 法

Sasano らの提案している手法 [4] の問題点は，オノマトペではない文字列までも辞書に追加してしまう点であった．この問題を解決するためには，オノマトペではない文字列を辞書から取り除く必要がある．オノマトペパタンの文字列がオノマトペであるかどうかを決める特徴として，制約がある．また，このオノマトペパタンに課されてる制約は，音象徴に基づいているため，音象徴を考慮することは問題解決に有効であると考える．

このため，本研究ではオノマトペパタンだけでなく音象徴も考慮したオノマトペ判別モデルによる辞書の自動構築法を提案する．図 1 は，オノマトペ辞書構築方法の概要であり，処理過程の順に説明を行う．

3.1　候補語の前処理

まず，オノマトペパタンと一致する文字列を生成する．この際，使用するオノマトペパタンは，Sasano らの手法と比較を行うため，同じパタンのみを使用する．表 1 は，Sasano らが用いたパタンと具体例を示しており，S_1 と S_2 には，表 2 が示す文字列が当てはまる．また 2 文字から 4 文字の繰返しもオノマトペパタンとして用いている．

次に，生成した文字列を音素表記に変換し，音素単位に分割する．

3.2　phoneme2vec

この手順では，分割された音素をそれぞれ音素の意味を表す分散表現へと変換する．これは，分散表現を判別モデルに用いることで音素の意味を考慮するためである．本研究では，音素の意味を考慮した分散表現を構築するため，言語学者が音素の意味を定義する際に行った方法を模倣した phoneme2vec を提案する．

音素の意味を扱うには，2.4 節で述べたように，オノマトペを構成する音素の位置に注目する必要がある．文字列の位置関係から意味を扱う手法として word2vec がある [10,11]．word2vec とは，文字の意味を表す分散表現を構築する方法であり，厳密には文字の位置を考慮しない cbow と文字の位置を考慮できる skip-gram に分けられる．しかし，word2vec では，語を最小単位としているため，音素の分散表現は得られないため，本研究では，最小単位を音素とした skip-gram を用いて分散表現を構築する方法を採用する．また，分散表現に変換を行うことは語単位ではなく音素単位で可能であることを明示的にするため phoneme2vec と呼称する．

分散表現の構築には，NINJAL-LWP for BCCWJ（以後，NLB）[1]のオノマトペデータのうち，Sasano らが用いたパタンと一致する語を使用する．NLB とは，現代日本語書き言葉均衡コーパス [12] のオンライン検索システムであり，擬音語・擬態語 4500 日本語オノマトペ辞典 [13]，Dictionary of Iconic Expressions in Japanese [14]，Unidic オノマトペ辞書 [2]の三つの文献で述べられているオノマトペがまとめられている．音素の意味を表す分散表現の構築をオノマトペのみで行う理由は，オノマトペではない語の音素には音象徴を示さない場合があるためである [8]．

構築した分散表現を評価する際，人手で各語の類似度をまとめたデータセットを用いて，どれくらい同じように意味を表現できているかを確認する．しかし，各子音・母音の類似度をまとめたデータセットは存在せず，構築した分散表現が意味を正しく表現できているかどうかの判断はできない．このため，パラメタチューニングを行い，判別モデルの精度が高くなるよう設定した．また，パラメタの一つである窓枠は，候補語を構成する音素の位置を把握しなければいけないため，分散表現構築に使用するデータの最大長である 12 に設定した．

3.3　オノマトペ判別モデル

最終手順では，分散表現に変換された候補語がオノマトペであるかどうかを判別モデルで確認し，オノマトペであると判別された候補語のみを辞書にまとめる．ここで使用するオノマトペ判別モデルは，音象徴を考慮する必要がある．音象徴を考慮するためには，子音・母音が表す意味だけでなく，位置および共起の関係も考慮する必要がある．そこで本研究では，リカレントニューラルネットワーク（RNN）をオノマトペ判別モデルに採用する．これは RNN が，分散表現を用いて，時系列性のあるデータの正誤を判別することが可能だからである．つまり，RNN は音素の意味を表す分散表現を利用することが可能であり，文字列の位置関係を考慮することができるためである．また RNN に類似する手法として，LSTM や GRU が存在するがこれらの手法は RNN の時系列データが長すぎる場合に生じる問題を解決するために提案された手法であり，本研究で扱うデータの長さは最大 12 と短いため，RNN を採用した．

1：NINJAL-LWP for BCCWJ: http://nlb.ninjal.ac.jp（2020 年 03 月 19 日 閲覧）

2：UniDic: https://unidic.ninjal.ac.jp/（2020 年 03 月 19 日 閲覧）

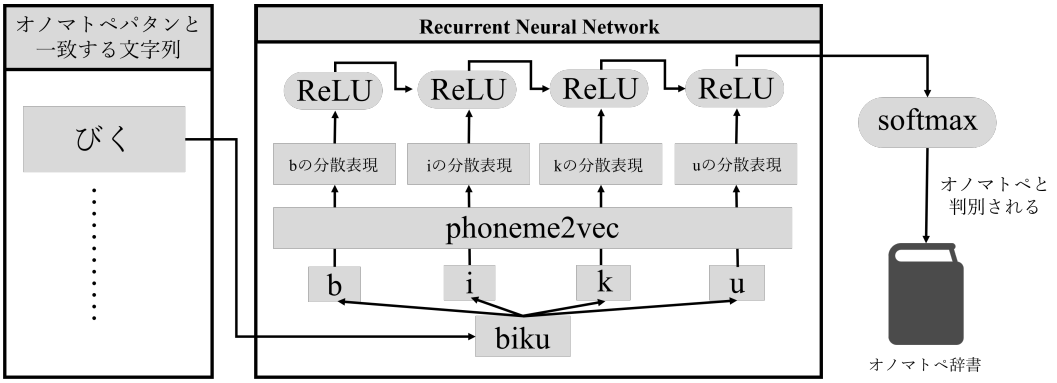

図 1　オノマトペ辞書構築の概要

表 3　オノマトペ判別モデルの判別精度りの良さ

		テストデータ	
		正	誤
提案手法	正	139.0（TP）	24.0（FP）
	誤	28.0（FN）	103.0（TN）

RNN にもハイパパラメタがあるため，パラメタチューニングを行う．今回行うオノマトペかどうかの判別は二値分類としてとらえることが可能であるため，損失関数には交差エントロピーを採用する．また残りのハイパパラメタである活性化関数には，中間層では学習時間の向上が可能であり，深層学習にて標準的に使用されている ReLU 関数 [15] を，出力層では softmax 関数を採用し，オプティマイザには，現状標準的もしくは最適であるとされているものはないため，調べた限り最も新しいとされている手法である AMSGrad [16] を採用する．

オノマトペ判別モデルの学習データには，正例として NLB のオノマトペデータを，負例として日本国語大辞典 [17] の見出し語のうち，正例と一致しない語を使用する．日本国語大辞典を負例データに採用した理由は，最も古くから存在する辞典の一つであり，オノマトペではない文字列を網羅的に扱うことが可能なためである．また既存手法と公平な比較を行うため，学習データのうち，表 1 のオノマトペパタンと一致するデータのみを判別モデルの学習に用いる．また，判別モデルの学習は層化 10 分割交差検証を行う．これは，過学習を防ぎ，汎用的なオノマトペ判別モデルの構築を目指すためである．

表 3 は，パラメタチューニング及び層化 10 分割交差検証を行い，最も良い精度を出したオノマトペ判別モデルの結果を表す混同行列である．表内の値はすべて，層化 10 分割交差検証を行い，得られた 10 回分の結果の平均値である．評価指標には，精度（Accuracy, A_{cc}），真陽性率（True Positive Rate, TPR），真陰性率（True Negative Rate, TNR）を用いる．精度とは，テストデータに対するモデルの正誤判別の正答率であり，式 (1) から求める．

$$A_{cc} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

真陽性率とは，テストデータ内のオノマトペであるデータに対する正答率であり，式 (2) から求める．

$$TPR = \frac{TP}{TP + FP} \tag{2}$$

テストデータ内のオノマトペではないデータに対する正答率である真陰性率を式 (3) から使用する．

$$TNR = \frac{TN}{TN + FN} \tag{3}$$

結果，表 3 より，精度は，0.819，真陽性率は，0.832，真陰性率は，0.811 であることが分かった．この結果より，提案手法のオノマトペ判別モデルは既存手法がオノマトペとして判別しているオノマトペパタンと一致する文字列のうち，実際にオノマトペである文字列の 81.9%を正しくオノマトペであると判断し，実際にはオノマトペではない文字列の 83.2%をオノマトペではないと正しく判別できたことが分かる．

4　評 価 実 験

本研究の目的は，オノマトペ辞書の中からオノマトペではない文字列を取り除き，オノマトペである文字列のみが含まれている質の良い辞書を構築することである．このため，Sasano らの手法 [4] と提案手法の比較を行い，構築した我々の辞書の質を確認する．

4.1　構築した辞書の評価方法

構築したオノマトペ辞書の有用性を評価するため，商品の使用感について記述されるレビューや SNS に対する精度を既存手法と比較する必要がある．このため，まず，Sasano らの手法 [4] が評判分析の対象であるドメインの文においてもよい精度を示すかを確認する．具体的には，Sasano らの手法 [4] より構築されたオノマトペ辞書を用いた形態素解析の結果を人手で確認し，オノマトペであると判別された形態素に対する精度を確認する．この精度は，形態素解析の結果より，オノマトペであると判断された形態素の内，正しく判別された形態素の割合を意味する．精度を確認する文には，Twitter 社の SNS [3]から公開されているツイート文を利用する．これは，評判分析を行う際にしばしば用いられる文であり，Sasano らの手法 [4] では精度が確認されていないドメインの文であるためである．本来

3 : Twitter: https://twitter.com/（2020 年 03 月 19 日　閲覧）

表 4　Sasano ら [4] の手法のオノマトペ判別精度

	正	語
既存手法	233	167

表 5　ツイートデータに対するオノマトペ判別モデルの結果

		テストデータ	
		正	誤
提案手法	正	222.0（TP）	61.0（FP）
	誤	11.0（FN）	106.0（TN）

であれば，既存手法が構築した辞書を用いた形態素解析の結果，オノマトペであると判別された形態素全てを確認し，精度を求めることが理想であるが，全ツイート文を使用することは不可能である．このため，オノマトペであると判別された形態素数を表す標本サイズは無作為検出における母比率と標本比率の許容標準誤差の下記計算式 (4) から算出された件数とする [18]．

$$\sigma_p \geqq Z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{4}$$

式 (4) 中の p は母比率であり，$\hat{p}$ は標本比率を意味するまた σ_p は比率の標準誤差を，Z_α は有意水準 α % における標準正規分布に従う統計量 Z を表しており，n は標本サイズである．また標本比率 $\hat{p}$ が不明なため，標本誤差が最小限になるよう標本サイズを最大値にする標本比率 $\hat{p} = 0.5$ に設定したところ，標本サイズ n は式 (5) のように求まったため，この件数を超えた 400 件と設定した．

$$n \geqq Z_\alpha^2 \frac{\hat{p}(1-\hat{p})}{\sigma_p^2} = 384.16 \fallingdotseq 400 \tag{5}$$

ツイート文の取得には，リアルタイムに投稿されているツイート文をランダムに取得することが可能な，Twitter 社が提供する API である Sample realtime Tweets [4] を使用する．この方法で集めた標本は，ツイート文に含まれるオノマトペ母集団の特徴を検出した縮図のようなデータ集合であるといえるため，母集団にみられる特徴と同様の特徴を表すことが保証される．

Sasano らの手法 [4] より得られた結果を確認ところ，表 4 が示す結果が得られた．この表より，オノマトペであると判別された文字列のうち，正しくオノマトペであると判断されていた割合は 57%と低い値であることが分かる．また，この原因は，オノマトペと判別されたオノマトペではない形態素が誤って辞書に登録されているためであることを確認した．つまり，オノマトペパタンで表される文字列には，オノマトペではない文字列が含まれているため，この規則の曖昧性がオノマトペ辞書構築に悪影響を与えていることが分かった．

次に 3.3 節で得られたオノマトペ判別モデルの有用性を評価するために，Sasano らの手法がオノマトペであると判別した文字列 400 件に対する精度を確認する．表 5 は提案手法の結果であり，提案手法の真陽性率は，95.7%，真陰性率は，62.6%であった．この結果より，Sasano らの手法がもつ問題点である，辞書に登録されているオノマトペではない文字列を 62.6%の精度で辞書から取り除くことができ，オノマトペである文字列を 95.7%の精度で辞書に登録することが可能であることが分かった．以上より，提案手法は Sasano らの手法 [4] の問題点を改善し，有効性を示したといえる．

4.2 考　　察

今回構築したオノマトペ判別モデルのうち，最も精度が良いモデルで使用されていた phoneme2vec で構築した分散表現の考察を行う．Hamano により，似ているとされている子音 20 種 [8] について確認を行った．

結果，この分散表現により，意味に類似性がみられる音象徴をもつ 20 種類の子音のうち，16 種類の子音で似ているとされている子音との類似性が最も高いという結果が得られたこのため，おおむね音象徴を考慮した分散表現が構築できたと考える．しかし，今回構築した分散表現では，“ t ”，“ d”，“w ”，“ 子音なし ” の類似性を考慮することができていない．

今回構築したオノマトペ判別モデル誤って判別したオノマトペパタンと一致する文字列を集計し，図 2 のような積立棒グラフを作成した．横軸は，子音の種類を示し，縦軸は，各子音の出現頻度である．また色と対応している本例内の数字は，生成されたオノマトペパタンと一致する文字列内で何文字目であるかを意味している．図 2 より，“子音なし” が誤って判別された文字列での出現頻度が他の子音と比べて多いことが分かった．この理由としては，“子音なし” の表す意味を今回 phoneme2vec で構築した分散表現では扱えていないことが示されていることから，音象徴をうまく扱えていないことがオノマトペ判別モデルの精度低下に繋がったと考えられる．

本研究では，提案手法の比較を行うため既存手法で用いられていたオノマトペパタンのみを使用したが，それ以外にもオノマトペパタンは存在するため全オノマトペパタンを用いることで更に音象徴を考慮した分散表現の構築が可能であると考えられる．

5　おわりに

本研究では，商品に対する評価が記述されたテキスト文がしばしばみられる SNS や EC サイトの Web ページからデジタルマーケティングのような評判分析を行う際に有用だとされているオノマトペを検出するための辞書構築法を提案した．オノマトペ判別モデルには，音素の表す意味，位置および共起関係を同時に扱うことが可能である RNN を採用した．また音素の意味を扱うため，word2vec の音素単位である，phoneme2vec を用いて音素の分散表現を構築した．

評価実験では，既存手法で構築した辞書と提案手法で構築した辞書の比較を行った．この結果，提案手法によりオノマトペではない文字列を 62.6%の精度で辞書から取り除き，オノマトペである文字列を 95.7%の精度で辞書に登録できることが分かったため，既存研究のオノマトペではない文字列が辞書に含

4 : Sample realtime Tweets: https://developer.twitter.com/en/docs/tweets/sample-realtime/overview （2020 年 03 月 19 日　閲覧）

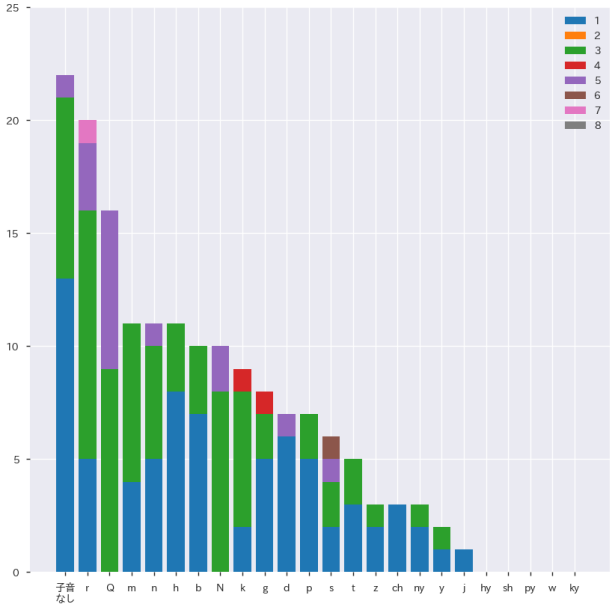

図 2　誤って判別された文字列の位置別使用頻度

まれてしまう問題点を改善できた．phoneme2vec で構築した分散表現が各子音・母音の音象徴を正しく表せているのかどうかを確認するため，他の音素と類似性を持つ 20 種類の音素でコサイン類似度を確認したところ，16 種類の音素で類似性のある音素とのコサイン類似度が最も高いことが分かり，音象徴を考慮した分散表現を構築できたと考える．また，phoneme2vec にて音素の意味を表す分散表現を構築できたことにより，オノマトペ間の意味の類似性を確認できるようになったと考える．

既存手法との比較を行うため，扱うオノマトペパタンを本研究では制限したが，実際は既存手法では用いられていないオノマトペパタンと一致するオノマトペも存在するため，それらも考慮したオノマトペ判別モデルを構築することで辞書の実用性を高める．

謝　　辞

本研究の一部は文部科学省私立大学戦略的研究基盤形成支援事業，JSPS 科研費 JP19H01138 の助成および同志社大学大学院文化情報学研究科の研究推進補助金を受けて遂行された．ここに記して謝意を表す．

文　　献

[1] 田守育啓, ローレンススコウラップ. オノマトペ —形態と意味—, 第 6 巻. くろしお出版, 1999.

[2] Fumiaki Saitoh, Hikaru Aoki, and Shohei Ishizu. Knowledge Extraction from Web Reviews Using Feature Selection Based on Onomatopoeia. In *Proceedings of HCI International 2015 - Posters' Extended Abstracts*, pp. 650–655. Springer International Publishing, 2015.

[3] Manabu Okumura, Atsushi Okumura, and Suguru Saito. Automatic Construction of a Japanese Onomatopoeic Dictionary Using Text Data on the WWW. In *International Conference on Application of Natural Language to Information Systems*, pp. 209–215. Springer, 2006.

[4] Ryohei Sasano, Sadao Kurohashi, and Manabu Okumura. A Simple Approach to Unknown Word Processing in Japanese Morphological Analysis. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 162–170. AFNLP, 2013.

[5] 筧壽雄, 田守育啓. オノマトピア 擬音・擬態語の楽園. 剄草書房, 1993.

[6] Ruslan Mitkov. *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2012.

[7] Shoko Hamano. *The Sound-Symbolic System of Japanese*. CSLI Publications,Kurosio, 1998.

[8] 浜野祥子. 日本語のオノマトペ: 音象徴と構造. くろしお出版, 2014.

[9] Noriko Iwasaki, Peter Sells, and Kimi Akita. *The grammar of Japanese Mimetics: Perspectives from structure, acquisition, and translation*. Routledge, 2016.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pp. 3111–3119, 2013.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *Workshop Track Proceedings of 1st International Conference on Learning Representations*, 2013.

[12] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. *Language resources and evaluation*, Vol. 48, No. 2, pp. 345–371, 2014.

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

[14] Ikuhiro Tamori Hisao Kakehi, Lawrence Clifford Schourup. *Dictionary of iconic expressions in Japanese*. Trends in linguistics. Mouton de Gruyter, 1996.

[15] Hidenori Ide and Takio Kurita. Improvement of learning for cnn with relu activation by sparse regularization. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2684–2691. IEEE, 2017.

[16] Sashank Reddi, Satyen Kale, and Sanjiv Kumar. On the Convergence of Adam and Beyond. In *Proceedings of 6th International Conference on Learning Representations*, 2018.

[17] 日本語大辞典刊行会. 日本国語大辞典, 第 2 巻. 小学館, 2002.

[18] Cochran William, G. *Sampling techniques*, Vol. 2. John Wiley and Sons, Inc, 1963.

付　　録

1　図　　表

本章では，本文上に記載しなかった言語学者によって報告されているオノマトペパタンと音象徴についてまとめている．

表 A·1　1 モーラのオノマトペパタン

オノマトペパタン	具体例
CV	hu
CVQ	huQ
CVN	baN
CVV	gaa
CVVQ	baaQ
CVVN	baaN
上記パタンの繰り返し	huhu, huQhuQ, huhuhu, etc.

表 A·2　2 モーラのオノマトペパタン

オノマトペパタン	具体例
$CVCV$	gaba
$CVCVQ$	bataQ
$CVCVVQ$	baraaQ
$CVCVri$	batari
$CVQCVri$	baQsari
$CVCVN$	bataN
$CVQCVN$	baQtan
$CVCVVN$	bataaN
$CVQCV$	doQka
$CVNCV$	muNzu
$CVNCVri$	boNyari
$CVCVCVCVQ$	basbasaQ
$CVCVCVCVVQ$	basbasaaQ
$CVCVCVCVN$	pakapakaN
$CVCVCVCVVN$	pakapakaaN
$p_1(CVCV)p_2(CVCV)$	dotabata
$p_1(CVCVri)p_2(CVCVri)$	norarikurari
$p_1(CVCVN)p_2(CVCVN)$	gatoNgotoN
上記パタンの繰り返し	gabagaba, batabatabata, etc.

表 A·3　文献 [7,8] の定義するオノマトペパタンの V が持つ音象徴

	オノマトペパタンの V
a	広い，平ら，後半，目立つ
i	線状，細さ，高音，緊張
u	突き出る
e	野卑
o	目立たない

表 A·4　文献 [7,8] の定義する CV パタンの C が持つ音象徴

	CV の C	
p	破裂，破れる，完全に覆われる	軽い，小さい，細かい
b	膨張，肥満	重い，大きい，粗い
t	張力の弱い表面を叩く	軽い，小さい，細かい
d		重い，大きい，粗い
k	固い表面，動作の厳しさ，きつさ	軽い，小さい，細かい
g	確実さ，空洞から外への運動	重い，大きい，粗い
s	抵抗のない表面を滑る	軽い，小さい，細かい
z	液体，流動体	重い，大きい，粗い
n	捉えにくい	
y		
h	息	
w/子音なし	興奮，動揺	
m	抑圧	
r		
口蓋化	子供っぽさ，雑多なもの，制御の不十分さ	

表 A·5　文献 [7,8] の定義する $C_1V_1C_2V_2$ パタンの C_1 が持つ音象徴

	$C_1V_1C_2V_2$ の C_2	
p	張力のある表面，肥満	軽い，小さい，細かい
b		重い，大きい，粗い
t	張力の弱い表面，弛緩	軽い，小さい，細かい
d	目立たない	重い，大きい，粗い
k	固い表面，きつさ	軽い，小さい，細かい
g	確実さ	重い，大きい，粗い
s	抵抗のない表面，液体	静か，軽い，小さい，細かい
z	流動体	重い，大きい，粗い
n	滑り，捉えにくい，粘着性，のろさ	
y	揺れ，頼りない動き	
h	美しさ，弱さ	
w/子音なし	興奮，動揺	
m	抑圧	
r		
口蓋化	子供っぽさ，雑多なもの，制御の不十分さ	

表 A·6　文献 [7,8] の定義する $C_1V_1C_2V_2$ パタンの C_2 が持つ音象徴

	$C_1V_1C_2V_2$ の C_1
p	破裂，破れる，完全に覆われる，膨張，肥満
b	
t	叩く，接触，密着，合致
d	
k	開く，中から出てくる，上下か内外の運動
g	
s	接触しながら動く，摩擦
z	
n	力のなさ，折れ曲がる
y	輪郭がはっきりしない
h	
w/子音なし	柔らかさ，弱さ
m	
r	流れるような運動
口蓋化	子供っぽさ，雑多なもの，制御の不十分さ