CE-002

日本語の音韻変化規則に基づく未知語処理法の提案
An Approach for Unknown Word Processing based on Japanese Phonological Rules

馬場 睦也[1)]　楠 和馬[1)]　波多野 賢治[2)]
Tokiya Baba　Kazuma Kusu　Kenji Hatano

1　はじめに

テキスト検索やテキスト分類の研究を行う際に，形態素解析器は必要不可欠なツールとなっている．形態素解析とは，日常生活で使用している自然言語を，対象言語の文法や単語の品詞情報等に基づき，形態素と呼ばれる言語で意味を持つ最小単位に分割し，それぞれの形態素の品詞等を判別する処理のことを指す．日本語の形態素解析器として有名なツールには，JUMAN[1)] や MeCab[2)] などがあるが，そのバージョンアップに伴い，形態素として正しく判定される語 (以下，既知語) が増える反面，それぞれの形態素解析器が有する制約により，解析誤りが発生する場合がある．解析誤りは，ネットスラングや砕けた表現などが多い，ブログ記事やツイートなどに多く見られ，近年のネットユーザの動向把握の研究を精緻に行う上では大きな問題となっている．

解析誤りが発生する要因の一つに形態素解析器が使用する辞書 (以下，形態素辞書) に含まれない語 (以下，未知語) の存在がある．形態素解析時における未知語は，

- 既知語から派生した未知語
- 既知語と直接関連を持たない純粋な未知語

に分けられると言われており [1]，これらに対し，事前に形態素辞書に登録する語を増やしたり，形態素解析時に未知語を既知語として推定したりする網羅的な未知語処理が行われている [2, 3]．それでも，未知語の種類によってはそれぞれ個別ケースの対応求められ，更なる未知語処理を行うための対処法の一般化が困難であるため，文献 [1] で提案されているような人により観察され得る傾向を元に構築されたあるルールに従って，個別対応する手法を採らざるを得ない．

しかしながらこのようなルールベースの手法は，先にも述べたように何らかの方法であらかじめルールを設定する必要がある．ルール設定の際によく用いられるのは，そのほとんどが人による観察や計算機が発見したパタンとなることが多いが，そうした手法では，一般的に前者は見落とし，後者は発見されたパタン自体の重要性を一つ一つ人により確認出来ないことから生じる信頼性の欠落といった問題が起こり得る．

そのため本稿では，ルール設定の際の問題点を解決するために，文献 [1] がルール設定の際に着目していた語の表記やオノマトペが持つ特徴を参考にした結果，言語音の機能面に着目して抽象化を行う音韻論に基づくルール構築を行うことで，前述したルール設定の際の問題を解決し，形態素解析時に未知語と判定された語をオノマトペ生成パタンや音韻変化規則に基づき既知語と判定させる試みの結果報告を行う．言語学の一分野である音韻論に基づいたルールを構築するため，特に人による観察や計算機によるパタン発見に基づくルール設定に比べ，見落としなどの問題を回避出来るばかりか，言語学の一学問分野に裏付けられることによる信頼性の確保も実現できることから，既存のルールベースの手法に比べ更なる未知語処理を実行できる可能性が高まるものと思われる．

2　先行研究

1 節でも述べたように，未知語処理に関する研究は基本的に大規模コーパスから形態素辞書に登録されていない語を見つけ，それらを随時登録していくか，形態素辞書に登録されていない語を未登録語と判定し，それらの語の切れ目や品詞や語義を統計学的に推定した上で既知語と判定する方法 (推定できた確証が得られない場合は未知語と判定) が採られることが多い．

したがってルールベースの手法は，もっぱら特定分野，例えばツイートに対する未知語処理のような，ネットスラングや砕けた表現のような形態素辞書に登録されていない語が頻出するケースに特化した形で研究が行われていることが多い．こうしたルールは Project Next NLP[3)] などでも行われているエラー分析結果 [4] に基づいて構築されることも多く，1 節で挙げた文献 [1] も独自のエラー分析結果を使用した一例である．

文献 [1] では，分析対象は検索エンジンで使用されている Web ページとなっているが，未知語処理の対象には砕けた表現の代表格でもある長音化や小文字化といった未知語の種類が挙げられており，そうした種類の未知語に対する処理ルールを設定し，未知語処理を正しく行えるような工夫を行っている．一方，ツイートで扱われている未知語の種類には，新語・低頻度語や形態素辞書に未登録の固有名詞，表記揺れがほとんどであり，先の事例で挙げられていた長音化や小文字化といった未知語は低頻度であったことが報告されている [4]．こうした分析結果によって得られた知見をツイート自体にアノテーションし，それをコーパスとして用いる事で特に解析誤りが起きやすいツイートに対する形態素解析時の誤り解析を防ごうとする取組みも存在している [5]．

これらの既存研究は，いわばルール構築を人手で行うものであり，人手で観察されなかったものに対してはルール化されないといった問題を引き起こしかねない．そのため，こうしたルールのバリエーションを如何に拡張していくかという問題は，より高度な未知語処理においては特に重要な課題となる．そうした観点から行われている研究は，我々の知る限り文献 [6] しか存在しないが，この研究も基本ルールの設定は文献 [1] で設定されたルールに基づいており，基本ルールの設定時点での完成度が低い場合は，バリエーションの拡張に繋がらない可能性は否めない．

1）　同志社大学大学院文化情報学研究科
2）　同志社大学文化情報学部

1）　日本語形態素解析システム JUMAN: http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN.

2）　MeCab: Yet Another Part-of-Speech and Morphological Analyzer: http://taku910.github.io/mecab/.

3）　Project Next NLP: https://sites.google.com/site/projectnextnlp/.

表 1　1 モーラのオノマトペの生成パタン

パタン	例
CV	ふ
CV と	ふと
CVQ	ふっ
CVN	ばん
CVV	がー
CVVQ	ばーっ
CVVN	ばーん
CVQCVQ	くっくっ
CVNCVN	ばんばん
CVVCVV	がーがー

本稿の提案は，そうした基本ルール設定時点での完成度をより高く設定するために，言語音の機能面に着目して抽象化を行う音韻論に基づいたアプローチを採用したものであり，ルール設定の抽象化を高めることで構築された基本ルールの精緻化を図るものである．

3　提案手法

本節では，音韻論に基づいた未知語処理法について述べる．音韻論は，言語を音韻と呼ばれる，語の意味を変化させる働きを持つ音の最小単位で記述できるようにし，音の機能や構造，パタンを分析する理論，と言われている [7]．長年にかけて言語学分野において音韻に関する多様な定義がされてきており，音韻が変化しても語としての意味が変わらないパタンが数多く規則化されている．また，時代とともに派生語や造語が多く発生するといわれており [8]，オノマトペに対しても母音，子音，撥音，促音，長音などの音素を組み合わせた規則化もなされている [9, 10]．

そこで本稿で提案する未知語処理法では，音韻論に基づき一般化されたオノマトペの生成パタンや日本語の音韻変化規則を適用する．これにより，未知語処理規則の信頼性や網羅性を担保できる未知語処理法を目指す．

3.1　音韻に着目したオノマトペの生成パタン

音韻論では言語音を音素や音節などといった単位で表記することや，言語音を音の長さといった概念であるモーラを扱う [11]．日本語オノマトペの音韻形態は 1 もしくは 2 モーラの基本形にまとめることが可能であり，各モーラ数にはパタンがみられると述べられている [9]．例えば，“レート” は “レ”，“ー”，“ト” に分けられるため，3 モーラからなる単語である．1 モーラおよび 2 モーラのオノマトペの生成パタンを表 1，2 に示す．表 2 中の記号 C は子音，V は母音，Q は促音 (“っ”)，N は撥音 (“ん”) を意味している．p_1 と p_2 はそれぞれ丸括弧に入っている文字列が別の文字列であることを意味する．

3.2　音韻変化規則

本節では，自動的に既知語から派生した未知語を自動的に変形可能にするため，未知語処理における音韻変化規則を適用する．音韻変化規則にはさまざまな種類のものが挙げられているが，本稿の提案手法ではその中から規則が一般化され，処理の際に曖昧性が含まれていないものだけを採用する．

以下に挙げる四つの音韻変化規則が，本稿で用いた規則である．

- 母音融合

  母音融合とは，母音の連接を避けるために母音が

表 2　2 モーラのオノマトペの生成パタン

パタン	例
CVCV	がば
CVCVQ	ばた
CVCVri	ばたり
CVCVN	ばたん
CVQCV	どっか
CVNCV	むんず
CVQCVri	ばっさり
CVNCVri	ぼんやり
CVCV	ばさばさ
p_1(CVCV)p_2(CVCV)	どたばた
CVCVriCVCVri	ばたりばたり
CVCVNCVCVN	ばたんばたん
p_1(CVCVri)p_2(CVCVri)	のらりくらり
p_1(CVCVri)p_2(CVCVri)	がたんごとん

変化する規則である．窪園は，唇や舌などの調音器官の状態や音の特性など言語音の性質を表す素性である弁別素性を用いて語の表示を行えば，母音融合の一般化が可能である，と述べている [12]．弁別素性の組合せに基づいた母音融合の変化規則を表 3 に示す．

表 3　母音融合の規則

変化前		変化後	変化前		変化後
a + i	→	e	a + u	→	o
a + e	→	e	a + o	→	o
i + u	→	u	i + e	→	i
i + o	→	u	u + i	→	i
u + e	→	i	u + o	→	u
e + a	→	a	e + i	→	e
e + u	→	o	e + o	→	o
o + a	→	a	o + i	→	e
o + u	→	o	o + e	→	e

- 重音脱落

  重音脱落とは，同音の音節が連続している場合，一方が脱落する現象である [11]．例としては，表 4 に示すものが挙げられる．

表 4　重音脱落の例

変形前	変形後
ナガアメ（長雨）	ナガメ
ミチノオク（陸奥）	ミチノク
マツウラ（松浦）	マツラ

- 長音の短音化

  長音の単音化は，語末にある長音が短音化する現象のことである [11]．例としては，表 5 に示すものが挙げられる．

- 連続同音母音の長音化

  同音の母音が連続している場合，先頭の母音以降の母音が長音になる変化を連続同音母音の長音化という [11]．例としては，表 6 に示すものが挙げられる．

表 5　長音の短音化の例

変形前	変形後
オニンギョウ（お人形）	オニンギョ
シンコウ（新香）	シンコ
ビンボウ（貧乏）	ビンボ

表 6　連続同音母音の長音化の例

変形前	変形後
オカアサン（お母さん）	オカーサン
オニイサン（お兄さん）	オニーサン
オネエサン（お姉さん）	オネーサン

4　評価実験

本節では，本稿で提案した音韻論に基づく方法と文献 [1] で提案された方法による未知語処理性能の比較実験を行う．比較対象には，提案手法の適用前後の未知語の総数をカウントすることにした．また，未知語が既知語になった事例を目視し，その結果から考察を行う．

4.1　実験内容

本節では，実験の方法について説明する．本実験で扱う形態素解析器は，文献 [1] で提案されている未知語処理方法 (以下，既存手法) が実装された JUMAN Ver. 7.01 [13] を用いることとした．その理由は，既存手法により処理できなかった未知語を提案手法により既知語に判定できれば，提案手法の有用性を確認出来ると考えたからである．また，未知語処理の対象にするデータは，Twitter 社の SNS サービス[4]であるツイートデータを利用する．これは前述したように，ツイートデータには，新語や派生語が多用されており，砕けた表現がつぶやきとして投稿されることが多いため [14]，未知語処理の性能を評価するために妥当なデータだといえるからである．ツイートデータの取得には，リアルタイムに投稿されているツイートをランダムに取得可能な，Twitter 社が提供する API である Sample realtime Tweets[5] を用いる．

以上のツールを用いて，次の手順で未知語処理の実験を，既存手法と提案手法のそれぞれで行う．

1　API を介してツイートを 400 件取得する．
2　各ツイートを形態素解析器 JUMAN に入力し，未知語を得る．
3　(提案手法のみ) 取得した語に対して未知語処理を施す．
4　未知語処理済みの語を再び形態素解析器に入力し，既知語に変化したかを確認する．

なお，ランダムに取得するツイートの件数は，標本サイズは無作為抽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 (1) から算出した．

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

式 (1) 中の p は母比率，$\hat{p}$ は標本比率，σ_p は比率の標準誤差，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，n は標本サイズを表している．本実験では標本比率 $\hat{p}$ が不明なため，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．また，有意水準 α は一般的に使用される 5 % とし，比率の標準誤差は 5 % 以内とした．これらより，標本サイズ n は下記の式 (2) のように求まった．

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

作成した標本は，ツイート母集団の特徴を抽出した縮図のようなデータ集合であるため，母集団が有するデータの特徴を標本も有すると保証することができる．

4) Twitter: https://twitter.com/（2018 年 6 月 28 日閲覧）
5) Sample realtime Tweets: https://developer.twitter.com/en/docs/tweets/sample-realtime/overview（2018 年 6 月 28 日閲覧）

表 7　提案手法の適用効果

標本#	形態素総数	未知語総数 (既存手法)	未知語総数 (提案手法)	既知語判定率
1	8,546	95	61	0.358
2	8,950	86	49	0.430
3	9,221	107	65	0.393
4	8,821	80	58	0.275
5	9,476	116	76	0.345
6	8,741	97	67	0.309
7	9,001	119	72	0.395
8	9,642	108	63	0.417
9	8,787	81	45	0.444
10	8,659	95	44	0.436
平均	8984.4	96.7	60	0.380

4.2　実験結果

既存手法と提案手法の未知語総数をカウントした結果を表 7 に示す．表 7 の形態素総数は，標本ツイートに既存手法を用いて形態素解析を施した際に既知語と判定できた形態素数であり，未知語総数 (既存手法) は形態素総数のうち既存手法が未知語と判定した記号以外の語の総数である．一方，未知語総数 (提案手法) は既存手法を適用して未知語と判定された文字列に対して提案手法を適用し，JUMAN で再度形態素解析を行ったときに未知語と判定された記号以外の語の総数であり，既知語判定率は提案手法により既知語に判定できた語数を未知語総数 (既存手法) で割った値である．

ここで，表 7 で示した各手法の未知語総数が，統計的に有意な差があるかどうかを確かめるため，対応のある 2 標本 t 検定を行う．検定で取り扱う統計量 t は式 (3) により求める．

$$t = \frac{\bar{d} - \mu}{\sqrt{\frac{S^2}{n}}} \tag{3}$$

式 (3) 中の $\bar{d}$ は 未知語総数の差の平均，μ は母平均，S^2 は不偏分散，n は実験の試行回数を表している．検定の帰無仮説 H_0 は「提案手法と既存手法の未知語総数に有意な差があるとはいえない」，対立仮説 H_1 は「提案手法と既存手法の未知語総数に有意な差がある」とする．以上のように対応のある 2 標本 t 検定を行った結果，t 統計量は 15.727 となり，統計量に対する p 値は，0.01 を大きく下回った．したがって，有意水準 1% としたときに，既存手法と提案手法間に未知語総数に有意な差があった，つまり，提案手法の未知語総数のほうが有意に少ない，と結論付けることができる．

4.3　考察

未知語から既知語へと変化した語を確認したところ，“シィィッ” や “デュゥゥゥン” などのオノマトペが正し

く未知語処理されている一方，“ブチギレ” や “マジ” などのオノマトペではない語が誤ってオノマトペであると判断されている事例が確認できた．

また，音韻変化規則の重音脱落においても正しく処理がされている場合とされていない場合が確認され，正しく処理されている例として “ウーーーーーーパーーーーーーーールーーーーーーーーパーーーーーーー” を “ウーパールーパー” に，誤った処理の例と “アイナナ (ゲーム名: アイドリッシュセブンの略称)” を “アイナ” へと変換していることを確認した．一方，音韻変化規則の連続同音母音の長音化や母音融合においては，正しく処理がなされた事例が確認できず，“イイゾ” を “イゾ” と変換したり，“ヒラモトアイナ (人名: よしもとの芸人)” を “ヒラモチナ” へと変換していることを確認したが，音韻変化規則により変換された事例自体はオノマトペの生成パタンに比べると数が少なかった．

これらの結果より，人名を誤って既知語に変換してしまう可能性，“ブチギレ” などの既知語の一部を組み合わせた未知語を誤って既知語に変換してしまう可能性があることが判明した．つまり，今回発見された問題点は本稿の提案手法では対応ができないため，人名の場合には未知語の中に姓名で使用されている文字を使用することで，また，既知語の一部を組み合わせた未知語の場合には既知語の一部を組み合わせて対応することで，誤りを防ぐ必要がことができると考えられる．

また，未知語処理の各プロセスを行う順序によって，誤った変換が行われてしまっている可能性も考えられる．例えば，“ジィィィィィィイィ” という語に対して，重音脱落に基づいて変換した場合は “ジ” となるが，連続同音母音の長音化に基づいて変換した場合は，“ジー” となる．したがって，このような事例に対処できるよう，未知語処理の各プロセスの順序の最適化を図る必要がある．

5 おわりに

本稿では，言語音の機能面に着目して抽象化を行う音韻論に基づくルールを適用した未知語処理法を提案した．提案手法を用いることで，既存手法による未知語総数の約 38% を既知語へ変換することができる可能性がある．

しかしながら，現時点では提案手法による既知語への変換の正しさに関しては確認出来ていない状況である．そのため，今後の課題としては，提案手法で既知語に変換された語の正確性を確認すること，および文献 [1] で提案されているルールが適用されていない JUMAN Ver. 5.1 の出力結果に対し本手法を適用することで，提案手法の変換でどの程度の効果が得られるのかを確認すること，さらに，3 節で説明した未知語処理手順内の各プロセスを適用する適切な順序を実験的に精査する必要がある．

謝辞

本研究の一部は，日本学術振興会科学研究費補助金 (18H03342)，および北見工業大学工学部地域未来デザイン工学科テキスト情報処理とインフォマティクス研究室との共同研究によるものである．ここに記して謝意を表す．

参考文献

[1] 笹野遼平, 黒橋禎夫, 奥村学. 日本語形態素解析における未知語処理の一手法—既知語から派生した表記と未知オノマトペの処理—. 自然言語処理, Vol. 21, No. 6, pp. 1183–1205, 2014.

[2] Shinsuke Mori and Makoto Nagao. Word Extraction from Corpora and Its Part-of-speech Estimation Using Distributional Analysis. In *Proceedings of the 16th Conference on Computational Linguistics*, pp. 1119–1122. ACL, 1996.

[3] Masaaki Masaaki. A Part of Speech Estimation Method for Japanese Unknown Words Using a Statistical Model of Morphology and Context. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 277–284. ACL, 1999.

[4] 鍛治伸裕, 森伸介, 高橋文彦, 笹田鉄朗, 斉藤いつみ, 服部圭悟, 村脇有吾, 内海慶. 形態素解析のエラー分析. 言語処理学会第 21 回年次大会発表論文集, 2015.

[5] 大崎彩葉, 唐口翔平, 大迫拓矢, 佐々木俊哉, 北川善彬, 堺澤勇也, 小町守. Twitter 日本語形態素解析のためのコーパス構築. 言語処理学会第 22 回年次大会発表論文集, 2016.

[6] 斉藤いつみ, 貞光九月, 浅野久子, 松尾義博. 文字列正規化パタンの獲得と崩れ表記正規化に基づく日本語形態素解析. 自然言語処理, Vol. 24, No. 2, pp. 297–314, 2017.

[7] 立石浩一. 机上の音韻論: 音韻論者がするべきこと. 音声研究, Vol. 4, No. 3, pp. 40–43, 2000.

[8] 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, 第 2003 巻, pp. 63–70. IPSJ, March 2003.

[9] 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.

[10] 角岡賢一. 日本語オノマトペ語彙における形態的・音韻的体系性について. くろしお出版, 2007.

[11] 安部清哉, 加藤大鶴, 吉田雅子. 日本語の音. 朝倉書店, 2017.

[12] 窪園晴男. 日本語の音声. 岩波書店, 1999.

[13] 黒橋禎夫, 河原大輔. 日本語形態素解析システム JUMAN version 7.0, 2012.

[14] Suman Maity, Anshit Chaudhary, Shraman Kumar, Animesh Mukherjee, Chaitanya Sarda, Abhijeet Patil, and Akash Mondal. Wassup? lol: Characterizing out-of-vocabulary words in twitter. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, CSCW'16 Companion, pp. 341–344. ACM, 2016.