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

勝木 健太[†]　笹野 遼平[‡]　河原 大輔[§]　黒橋 禎夫[§]

katsuki@nlp.kuee.kyoto-u.ac.jp　sasano@pi.titech.ac.jp　{dk,kuro}@i.kyoto-u.ac.jp

[†]京都大学 工学部　[‡]東京工業大学 精密工学研究所　[§]京都大学 大学院情報学研究科

1　はじめに

Webの発展、およびそれに伴うブログ、ミニブログ、SNSなどのCGM (Consumer Generated Media) の一般化により、さまざまな情報がそれらのメディアから発信されるようになった。人々は、言語処理システムを通して、これらの情報を入手し利活用している。しかし、それらのメディアのテキスト中には、口語的表現、非正規表現、顔文字など、さまざまな言語表現バリエーションが溢れており、これが言語処理システムにおける主要なエラー原因の一つとなっている。つまり、言語処理システムの基盤技術である形態素解析において、それらの言語表現バリエーションが辞書に登録されていないために未知語となり、正しい形態素区切りが得られないなどの解析誤りを引き起こしている。

未知語は、表1のように分類することができる。これらの未知語への対処方法として、古くから人手で辞書エントリを追加するということが行われてきたが、明らかに限界がある。近年、コーパスから未知語を自動獲得する研究が行われており [1, 4, 5]、特に固有名詞や新語などopen classの語を獲得するためには適している。この手法の問題としては、低頻度の語を獲得することが難しいことや、未知語を解析できるようにするためには未知語獲得、辞書追加という二段階を適切なタイミングで行う必要があることが挙げられる。

既知形態素からの派生語やオノマトペについても、コーパスから獲得することが考えられるが、一方で、形態素解析において動的にこれらの未知語を推測する手法が考えられる。これは、既知形態素とのマッチングやパターンに基づいて未知語をオンラインで推測する手法であり、コーパスに基づく手法の欠点がなく、単独の事例から推定可能な場合に適した手法である。本論文では、これらの未知語のうち、小文字化・長音化による非正規表現と非反復型オノマトペを対象に、

表1: 未知語の分類 (下線はJUMANで扱っていることを示し、二重下線は本論文での対象を示す。)

既知形態素からの派生	(例)
表記ゆれ	素晴しい
連濁	(堀り) ごたつ
長音化	おいしーい
小文字化	ぁなた
記号化	ぁや ι い
口語的表現・方言	やっぱ
既知形態素からの派生以外	(例)
反復型オノマトペ	ほいほい
非反復型オノマトペ	ぺっちゃり
感動詞	いやっほー
顔文字・アスキーアート	(´・ω・`)
新語	tsuda る
固有名詞	Windows Azure

形態素解析の辞書引き時にこれらの可能性を動的に考慮する手法を提案する。なお、表記ゆれの認識については文献 [2]、連濁と反復型オノマトペの認識については文献 [6]、新語や固有名詞の自動獲得については文献 [5] を参照されたい。

2　形態素解析の概略と基本方針

まず、形態素解析の手順を簡単に述べる。形態素解析は通常、以下のような手順で行われる。

手順1 入力された文に対し、文中の各位置から始まる可能性のある形態素すべてを検索する。

手順2 形態素の候補を列挙したグラフ構造 (ラティス構造) を作成する。

手順3 形態素同士の組み合わせの中から、文として最も確からしい形態素の並びを決定する。

たとえば、「軽ーくはねる」という文が入力された場合、図1に示すラティス構造が作られ、最終的に太線で記されている組合せに決定される。点線部分は、提案手法によって追加されるパス (後述) である。

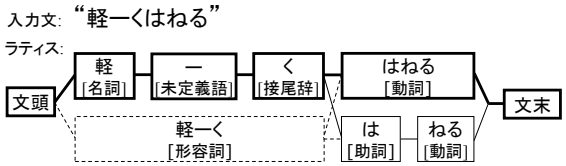

図 1: ラティス構造の例

本研究では形態素解析器として JUMAN[1] を用いる。JUMAN では、**手順 3** において形態素の並びを決定する際、人手で設定した連接コストや単語生起コストから、それぞれの形態素列のコストを計算し、もっともコストの小さい形態素の並びに決定する。ChaSen[2] や MeCab[3] など、機械学習を用いた形態素解析器においても基本的な解析の流れは同様である。

通常、**手順 1** では辞書に登録されている形態素が検索される。本研究では、長音化・小文字化した表記と非反復型オノマトペを認識するために、長音や小文字を置換、削除した文字列も辞書検索の候補とし、またオノマトペのパターンにマッチするものも形態素の候補として追加する。なお、これらの候補に対するコストは、解析誤りが少なくなるように経験的に設定する。このように本手法は、非正規表現を形態素解析の後処理で認識する手法 [7] とは考え方が異なっている。

3 非反復型オノマトペの自動認識

オノマトペとは「ほいほい」「ぺっちゃり」などのような擬音語・擬声語のことである。比較的自由に生成できることから辞書に載っていない語も多く存在し、形態素解析の誤り原因の一つとなっている。しかし、オノマトペの多くはいくつかのパターンで記述できることが知られており [3]、笹野らは形態素解析において「ほいほい」「こっくりこっくり」などのような 2〜4 文字の反復が出現した場合にそれらを形態素候補に加えることで、70%〜90%の精度で反復型オノマトペの自動認識できることを報告している [6]。

本研究ではさらに、表 2 に示すような特定のパターンに該当する文字列をオノマトペの可能性があるとし形態素の候補に加えることにより、反復を含まないオノマトペの自動認識を行う。オノマトペの取り得る品詞としては副詞、形容詞、サ変名詞などがあるが、本研究では主に形態素区切りの改善を目的とし、すべて副詞として扱う。表 2 中のコストは形態素解析器 JUMAN における単語生起コストを表している[4]。

表 2: 非反復型オノマトペのパターンとコスト

	パターン*	コスト	例
1	HっH り	300	ぽっこり
2	KッK リ	300	マッタリ
3	HっHY り	300	ぺっちゃり
4	KッKY リ	300	ポッチャリ
5	KKっと	200	サクっと
6	KKッと	200	パキッと

*H は平仮名、K は片仮名、Y はヤ行拗音字を表す

4 長音化表記の自動認識

長音化表記としては、長音による置換と長音の挿入の二種類がある。たとえば、「おいしい」に対して「おいしー」は「い」が長音に置換されており、「ぜんぶ」に対して「ぜーんぶ」は長音が挿入されている。これらは、特にブログなど感情を込めたテキストに多く見られる。

本研究では、形態素の辞書引き時に、入力文そのままの文字列に加え、入力文中の長音を「あ」「い」「う」「え」に置換した文字列と、長音を削除した文字列も辞書引きを行う候補に加える。ただし、長音を置換する条件として長音の直前が平仮名であること、長音を削除する条件として長音の直前が平仮名または漢字であることとする[5]。長音を置換する文字は、長音直前の文字がア行なら「あ」、イ行なら「い」、ウ行なら「う」、エ行なら「い」と「え」、オ行なら「う」というルールに従うとする。これらの条件とルールは、コーパス中の長音化表記を観察した結果に基づき定めた。

たとえば、入力文字列「軽ーく」に対して、長音を削除した「軽く」を辞書引きし、辞書にあれば候補としてラティス構造に登録する (図 1)。また、入力文字列「おいしー」に対しては、長音を削除した「おいし」に加えて、長音の直前文字「し」がイ行であるため「ー」を「い」に置換した「おいしい」を辞書引きする。

5 小文字表記の自動認識

Web 上のテキストには「ぁなた」、「ヵわいい」などのように非正規的な小文字を用いた表現が存在する。これらの表現は「あなた」などのように通常の表記であれば容易に解析できる形態素であった場合でも、従来の形態素解析器では小文字部分を未知語と判定されてしまい、形態素解析の誤り原因の一つとなっている。

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?日本語形態素解析システム JUMAN
[2] http://chasen-legacy.sourceforge.jp/
[3] http://mecab.sourceforge.net/
[4] JUMAN では、一般的な副詞、名詞には 100、アルファベットを除く未知語には 1000 というコストが与えられている。
[5] 片仮名語の末尾に長音がある「コンピューター」などは、「コンピュータ」のバリエーションとして表記ゆれの枠組みで別途扱う。

表 3: 非反復型オノマトペの自動認識の精度

適合率	再現率	F 値
93/100	94/100	0.935

表 4: 非反復型オノマトペの自動認識結果の例

正しく解析できた非反復型オノマトペ (抜粋)

- ホクホクというか、**ポックリ** した食感で、やさしい甘さです。
- 設定は簡単だし、**サクっと** ゲームもできました。
- 髪の毛はキンキンの金髪で **ぺっちゃり** ねかせてました…。

誤って非反復型オノマトペであると認識されたもの

- 近くの ロッテリ 屋までテクテクお出かけしてみる。
- マンションリフォーム、ペットリ フォーム、
- 普通にちただけでも満足なのに凄い贅沢 いってりゅ俺 ‥
- ハッタリ だろヤムヤム？ま、まだ堕ちるなんて …!?
- 新聞も一面にそのニュースでも ちっきり だった。
- ミナたんは、チェイサーのツアラーSに のっちょり ますっ!!
- ナツヒやアカホシが喰 らったり したらどうなるんだろ。

認識できなかった非反復型オノマトペ (抜粋)

- でも、気長に**まったり**やっていきたいです。
- たまには徒歩で**まったり**ドラクエ世界を旅してみると、
- ウケたのが体温計をおケツに**ぶっすり**やられた瞬間、

本研究では「あいうえおわカ」が出現した場合、それぞれ「あいうえおわか」に変換した上で辞書引きすることにより、これらの表記に対処する。

6 実験

本研究の実験では、形態素解析器として JUMAN 6.0 を用いた。実験で用いるコーパスとしては、検索エンジン基盤 TSUBAKI[6] で対象としている Web ページ 1 億件から 20 文字以上でかつ平仮名を含む文を無作為に抽出し使用した。以下では、これを TSUBAKI コーパスと呼ぶ。

6.1 非反復型オノマトペ認識の評価

非反復型オノマトペ認識の評価実験に先立ち、表 2 に示すパターンに適合する未知のオノマトペを含むコーパスの作成を行った。まず、TSUBAKI コーパスから表 2 に示すパターンに適合するテキストを正規表現を用いて抽出し、コーパスの先頭から順に JUMAN の辞書に含まれないものについて人手でオノマトペであるかどうかのチェックを行い、100 個の未知のオノマトペを含む文集合を作成した。100 個目の未知のオノマトペが出現したのは 38,900 文目であり、表 2 に示すパターンで認識できる未知のオノマトペは Web テキスト約 390 文に 1 つ出現することになる。

[6]http://tsubaki.ixnlp.nii.ac.jp/

表 5: 長音化認識の評価結果

	A	**B**	**C**	**D**	合計
適合率評価	50	24	25	1	100
再現率評価	56	26	18	-	100

作成したコーパスを用いた再現率の評価実験、および、TSUBAKI コーパスを対象とした自動認識結果 100 個の適合率の評価実験の結果を表 3 に示す。適合率は 93%、再現率は 94%であり、非常に高い精度でオノマトペの認識に成功した。表 4 に新たに正しく解析できるようになった例、および、誤ってオノマトペであると認識された例、認識できなかったオノマトペ例を示す[7]。誤ってオノマトペであると認識された 7 例はいずれも、オノマトペの自動認識を行わなかった場合も正しく解析できなかった形態素であり、解析結果が悪化したと言えるものは存在しなかった。また、認識できなかったオノマトペ 6 例中、5 例は「まったり」であり、これは動詞の連用形「待ったり」を優先してしまった結果である。

6.2 長音化表記認識の評価

長音化表記認識の評価実験に先立ち、長音化表記のコーパス作成を行った。まず、TSUBAKI コーパスから平仮名または漢字に後続する長音を含む文を抽出し、次に、JUMAN 辞書に含まれない、固有表現でない、片仮名語を平仮名で表記したものでない、という三点を満たす長音化表記 100 個を人手で抽出した。100 個目が出現したのは 8,233 文目であり、長音化表記は Web テキスト約 82 文に 1 つ出現することとなる。

このコーパスを用いた評価結果を表 5 の再現率評価に示す。また、表の適合率評価は、TSUBAKI コーパスの先頭から長音化認識を適用した 100 個を抽出し評価した結果である。この四つの分類は、長音化認識を行わない場合と比較し、解析結果を以下のように分類したものである[8]。

A 正しく解析できていなかった表現について、区切り、品詞ともに正しく解析
B 正しく解析できていなかった表現について、区切りは正しくなったが、品詞は誤って解析
C 正しく解析できていなかった表現を誤って解析
D 正しく解析できていた表現を誤って解析

表 6 に、それぞれの分類の解析例を示す。形態素解析結果としては **A** 以外は誤りであるが、**B** と **C** に分

[7]解析結果例において、下線 は提案手法により一つの形態素と認識された箇所を、**太字**は注目箇所における望ましい形態素を表す。
[8]再現率評価における長音化表記は、長音化認識を行わない場合はすべて解析誤りとなるため、**D** に該当する例はない。

表 6: 長音化表記の認識結果の例

A 襟ぐり、袖口に **軽ーく** ゴムが入っている感じです。
A その内容は今度詳しくは報告いたし **まーーーす**！
A でもその後の、デザートはケーキで **ぜーんぶ** 食べたよ。
A 調子乗るとすぐ飲みすぎて体痛めるって子供みたいで **かわいー**。
A **どー** にもならんモノだが、コレを今日という日の証とします
B それ自体は **まー** いいんですが、「？でしょうか」といった･･･
B その奥さんが **すげー** 怒って大変だったし ‥
C 翔ちゃんすぐＺＥＲＯに帰っちゃう **じゃーん**
C なんでも **いー**が、実際の試合は一体何時何分からやるんだよ！
D ルアフ「**…あー**もう！わかったよ！こうなったら柏餅食べ･･･

類された表現はもともと正しく解析できていなかった表現であるので、解析結果が悪化したとは言えない。**B** と **C** の原因としては、「すげー」や「じゃーん」のような口語的表記に起因することが多かった。正しく解析できていた表現が誤って解析された **D** に分類されるものは表 6 に示す 1 例のみであった。これは「編もう」に対する長音挿入と誤認識された結果である。このように誤認識された結果は 1 例しかなく、他の解析にほとんど悪影響を与えることなく長音化認識を行うことができたと言える。

再現率評価において、**C** に該当する 18 例のうち、4 例は長音化の認識に失敗した。この例を次に示す。

(1) んー、いや待てよ、まず前提としてＩＰフィルタをＷｉｋｉ側に持つべきかどうか考えよう。

この例では、感動詞「ん」が辞書にないために、それに対して長音を付加した表記の認識に失敗している。

6.3 小文字表記認識の評価

小文字表記認識では、基本的に対象の小文字を含むテキストであれば常に解析結果が変化することから、自動認識結果の評価のみを行った。TSUBAKI コーパスを対象とし、小文字表記認識を行うことにより解析結果に変化があった箇所を先頭から 100 箇所、評価した結果を表 7 に示す。解析結果が改善しなかった 23 例のうち、もともと正しく形態素解析できていた周辺の形態素に悪影響があったものを “悪化”、それ以外のものを “その他” に分類している。

100 個目の小文字表記が出現したのは 5,076 文目、そのうち改善したものが 77 例あることから、66 文に 1 つの割合で解析結果が改善したことになる。ただし、解析結果が改善した 77 例のうち「なぁ」「ねぇ」「まぁ」「あぁ」「さぁ」「ヵ月」が 71 例を占めており、これらの表記を形態素辞書に登録することで 9 割は対処できると言える。表 8 に、解析が改善した残りの 6 例、解析が悪化した 3 例、および、その他に分類したものの例を示す。

表 7: 小文字を大文字化することによる解析の変化

改善	悪化	その他	合計
77	3	20	100

表 8: 小文字表記認識結果の例

正しく解析できた小文字表記 (抜粋)

- 彼は心のそこで思ってることが、まだ **ぁなた** に言えてない
- ＤＡＬ３つの個性で同時デビュー☆ **おにぃちゃん** 大好き！
- **ばぁちゃん** の作る「イチジクの甘露煮」は
- 本人曰く、感動して泣きそうだった **らしぃ**。。。
- わしは５歳の時に七五三やらなかったの **かぃ**？
- まぁ晴れを祈るぢゃあこのへんでぇ〜**ばぃばぃ**

解析が悪化した小文字表記

- おい、藤原１１月末までに小説出して**くれ**ぃ！
- ちゃんと見せたからゆみ **ぃの** 布団も見せて
- **いや**ぁ、勝手なまねが多くてすまないな。

解析が改善しなかった小文字表記 (その他) (抜粋)

- 「おばちゃん、あたくしは飼い犬じゃね ぇん だ！」と
- 知らなかったんだから硬いこと言うなっつぅ の。

7 おわりに

本稿では、小文字化・長音化による非正規表現と非反復型オノマトペを対象に、形態素解析の辞書引き時にこれらの可能性を動的に考慮する手法を提案した。残された課題としては、「すげー」「またーり」などの口語的表記、「ぁやぃぃ」のような記号化表記などがあるが、同様の枠組みで扱う予定である。

参考文献

[1] Shinsuke Mori and Makoto Nagao. Word extraction from corpora and its part-of-speech estimation using distributional analysis. In *Proceedings of COLING1996*, pp. 1119–1122, 1996.

[2] 岡部浩司, 河原大輔, 黒橋禎夫. 代表表記による自然言語リソースの整備. 言語処理学会第 13 回年次大会, pp. 606–609, 2007.

[3] 筧寿雄, 田守育啓. オノマトピアー擬音・擬態語の楽園. 勁草書房, 1993.

[4] 鍛治伸裕, 福島健一, 喜連川優. 大規模ウェブテキストからの片仮名用言の自動獲得. 電子情報通信学会論文誌, Vol. J92-D, No. 3, pp. 293–300, 2009.

[5] 村脇有吾, 黒橋禎夫. 形態論的制約を用いたオンライン未知語獲得. 自然言語処理, Vol. 17, No. 1, pp. 55–75, 2010.

[6] 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, pp. 819–822, 2007.

[7] 池田和史, 柳原正, 松本一則, 滝嶋康弘. くだけた表現を修正するための教師なし学習方式の提案と評価. In *FIT2009*, pp. 13–18, 2009.