新しい語・崩れた表記の処理

基般

笹野 遼平[1]　鍛治 伸裕[2]
[1] 東京工業大学精密工学研究所　[2] 東京大学生産技術研究所

情報発信者の多様化

ブログ，ミニブログ，SNS（Social Networking Service）などのCGM（Consumer Generated Media）の一般化により，さまざまな情報がこれらのメディアから発信されるようになった．それに伴い，多様な発信者が書いたテキストを目にする機会が増えてきており，これらのテキストを分析することにより，消費者の生の評価やニーズを把握したり，通時的変化を分析することで流行を定量的に観察することが可能となった．

しかし，このような多様な発信者が書いたテキストには，「ググる」「婚活」といった辞書に載っていない新しい語や，「知らなぃ」「もしも～～～し」のような崩れた表記など，新聞記事を主な対象としてきた従来のテキスト処理技術では対応できない表現が多く出現する．本稿ではこれらの表現への対処の試みを紹介する．

現実世界のテキストと形態素解析

多くの自然言語処理アプリケーションでは最初に形態素解析と呼ばれる処理が行われる．形態素解析とは入力されたテキストを，意味を持つ最小の単位である形態素に分割する処理であり，一般に図-1に示すような手順で行われる．この際，手順①において作成される形態素の候補は，基本的に事前に準備した辞書に含まれる形態素から作成される．このため，辞書に含まれていない表現を含む文は正しく解析できない場合が多い．

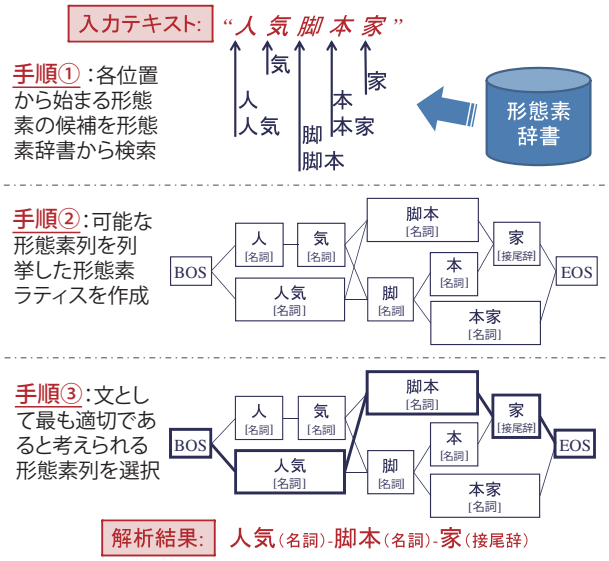

図-1　形態素解析の流れ

現在，一般的に使用されている形態素解析システムは新聞記事のような整ったテキストに対しては高い精度を実現している．しかし，「ググる」という新しい語や「知らなぃ」「ありがとー」などといった崩れた表記を含む図-2に示すようなWeb上のテキストに対しては十分な精度を実現しているとは言い難いのが現状であり，Webテキストに代表されるような現実世界のテキストに対応した言語処理技術が求められている．

新しい語の処理

我々が生活する社会においては，絶えず新しい語（以下では単に新語と呼ぶ）が作り出されている．たとえば「ググる」「婚活」などは，今から10年前には日本語に存在していなかった単語である．

図 -2　Web 上のテキスト（http://twitter.com より）

すでに述べたように，新語の大半は既存の辞書資源に含まれていないため，形態素解析における典型的な誤り要因の 1 つとなっている．従来の自然言語処理の研究においては，新聞記事のように新語が少ない領域が主要な研究対象になっていたという事情から，新語が深刻な問題として議論されることは稀であった．しかしながら，ブログや SNS などの CGM の拡大に伴い，この問題を無視できない状況が生まれつつある．

本稿では，新語を含んだテキストの頑健な形態素解析を実現するための技術として，分かち書きされていないテキストから新語を自動抽出することによって，新語辞書の構築を行う手法を紹介する．これは，単語が持つ文法的性質に着目することによって，新語抽出の問題を文字列の分類処理に落とし込むというものとなっている．

なお，以下では特に断りのない限り，本稿では片仮名語の新語を議論の対象とする．これは，新語の中に占める片仮名語の割合が高いことと，これまで筆者の 1 人が片仮名語の解析を中心に研究を行ってきたためである．ただし，以下で紹介する手法は片仮名語以外にも拡張可能であることを付け加えておく．

■基本アイディア

まず始めに，分かち書きされていないテキストから，辞書に登録されていない新語を自動抽出するための基本的な考え方について説明する．

ここでは例として，新動詞「ググる」を考える．今，分かち書きされていないテキストが手元にあり，そこから片仮名列「ググ」の出現個所を抽出したとする．そうすると，たとえば以下のようなテキスト断片が得られることが予想できる．

- 鯉のエサについてググってみる．
- ググらず分かったらマジすごいです．
- 分からないことはググれ．

これら片仮名列「ググ」の周辺には「ググる」が動詞であることを示唆する情報が豊富に存在する．たとえば，片仮名列「ググ」の直後には「って」「ら」「れ」などの文字列が見られるが，これらは動詞の活用語尾に特徴的な文字列である．さらに「ググら」の後には，否定を表す助動詞「ず」が出現している．このことから，片仮名列「ググら」は動詞の未然形である可能性が高いと推測できる．

この例から，ある文字列が単語であるかどうかを調べたい場合には，その文字列☆1 の周辺情報（主に後続文字列）を調べればよいことが分かる．さらに，周辺情報のパターンは，各品詞によって大きく異なるため，新語の品詞を推定することも可能である．たとえば，動詞の代わりに名詞の場合を考える．

- 検索エンジンはグーグルを使います．
- グーグルの T シャツを愛用する．
- 今日からグーグルで働きます．

名詞「グーグル」の後ろには「を」「の」「で」のような助詞が現れるなど，動詞の場合とは傾向が大きく異なることが確認できる．

■分類問題としての定式化

では次に，具体的にどのような計算手法を用いて新語抽出を実現するのかを説明する．

新語抽出というタスクは，次のようにすれば，分類問題に帰着させて考えることが可能である．まず，テキスト中に出現する全片仮名列☆2 を新語の候補と考える．そして，各片仮名列に対して，その文字列が単語であるか否かを判断する 2 値分類問題を解く．そうすると，単語であると判断されたにもか

☆1　動詞「ググる」のような活用語の場合は，語幹「ググ」を見ることになる．
☆2　実際には，計算効率のため，何らかのヒューリスティクスを使って候補を絞り込むことも考えられる．

かわらず，辞書に登録されていない片仮名列は新語であると考えることができるため，これによって新語抽出を行うことが可能となる．もし，新語の品詞推定を行いたいのであれば，2 値分類ではなく多値分類問題を扱えばよい．

こうした分類問題に対しては，機械学習のコミュニティにおいて研究されているさまざまな教師あり学習手法を適用することができる．実際に研究の中で使用したのは SVM（Support Vector Machine）である．ご存知の読者も多いだろうが，SVM は最大マージン原理に基づく分類器であり，高い分類性能を発揮することから，自然言語処理を含めた多くの分野において広く用いられている．

SVM に限らず，一般的に分類器の学習を行うためには，分類対象（ここでは片仮名列）を特徴量ベクトルに変換する必要がある．新語抽出を行う場合であれば，片仮名列の直後に出現する文字 *n*-gram[☆3] を特徴量として使うことができる．すなわち，各片仮名列はテキストに出現した文字 *n*-gram の異なり数と同じだけの次元数を持った特徴ベクトルへと変換される．特徴ベクトルの次元は個別の文字 *n*-gram に対応しており，今注目している片仮名列の直後にその文字 *n*-gram が出現していれば，対応するベクトルの要素は 1，そうでなければ 0 となる．このとき *n* の値は自由に定義できるが，実験では文字 1-gram から 5-gram までを用いた．

我々の実験において抽出された新語の一部を以下に示す（文献 1）からの抜粋）．

- コラボる，トイツる，ジコる，テソパる，デモる，タクる，ラチる，ヘチる
- イナタい，スンバラシい，ウッザい，ナヨい，ヘヴィい，ズブい

「コラボる」「トイツる」「イナタい」など，新語が抽出されていることが分かる．こうした新語は，従来の形態素解析技術では正しく解析することが困難であるとされていたが，このようにして構築された新語辞書を活用すれば，正しく解析することが可能になる．

☆3 文字 *n*-gram とは連続する *n* 文字の文字列のこと．

タイプ	例（括弧内は元の語）
長音記号の挿入*	でーす、もしも〜〜〜し
母音の挿入	やったあ、行けえええ
小書き文字の挿入*	見たぁぁい、ねむぅい（眠い）
促音・発音の挿入	すっばらしい、すんばらしい
長音記号による置換*	ありがとー、ねーさん（姉さん）
小書き文字による置換*	おぃしぃ（おいしい）、かわぃい（かわいい）
類似記号による置換	あやιい（怪しい）、こωばωわ（こんばんは）

表 -1 Web テキストに出現する崩れた表記（* は JUMAN7.0 で対応済みであることを表す）

崩れた表記の処理

Web テキストなど現実世界のテキストの処理を行う上で問題となる表現には，新語以外にも，「ぉぃしぃ」や「もしも〜〜〜し」などといった崩れた表記がある．ここからは，このような崩れた表記を含んだテキストの処理を実現するための取り組みの 1 つとして，形態素解析システム JUMAN7.0[2] における取り組みを紹介する．

■ Web テキストに出現する崩れた表記

まず表 -1 に，Web 上のテキストに出現する代表的な崩れた表記を示す．これらはいずれも辞書に登録されている語（以下では既知語と呼ぶ）に，長音記号や母音字，小書き文字が挿入されたり，一部の文字が小書き文字等に置換された表現であり，一般的な形態素解析システムでは正しく解析できない．

- ぉぃしかったでーーす

たとえば上記のようなテキストが入力された場合を考えると，「おいしい」（形容詞）や「です」（助動詞）という形態素が辞書に登録されていたとしても，「ぉぃしかった」や「でーーす」という表記に対応していないため正しく解析することができない．

形態素解析における辞書に含まれていない表現への対応策としては，前章までで紹介したように大量のコーパスから事前に新語や業界用語などといった未知語を獲得しておく手法と，統計情報や機械学習を用いて未知語モデルを学習する手法の 2 つがよく

タイプ	削除対象	削除する条件
長音記号の削除	ー、～	1. 直前が平仮名、または漢字 2. 一部の品詞（接頭辞、格助詞等）でない
小書き文字の削除	ぁ、ぃ、ぅ、ぇ、ぉ	直前の文字が平仮名で、かつ、削除対象がその平仮名を長音化させる文字である場合 (e.g. 削除対象が「ぁ」で、かつ、直前の文字が「か」)
タイプ	**置換対象**	**置換する文字**
長音記号の置換	ー、～	直前の文字が 1.「が」「ば」「ま」「ゃ」なら「あ」 2. い段、「え」「ね」以外の え段 なら「い」 3. う段 または お段 なら「う」 4.「え」「ね」なら「え」
小書き文字の置換	ぁ、ぃ、ぅ、ぇ、ぉ、ゎ、ヵ、	1.「ぁ」⇒「あ」2.「ぃ」⇒「い」3.「ぅ」⇒「う」 4.「ぇ」⇒「え」5.「ぉ」⇒「お」6.「ゎ」⇒「わ」 7.「ヵ」⇒「か」

表 -2　崩れた表現の正規化ルール

用いられる．しかし，これらの手法はいずれも，既知語からの派生ではない“完全な未知語”を扱う場合に適した手法であり，本稿で対象としている崩れた表記を扱うのに適した手法であるとは言えない．

まず，未知語を事前に獲得する手法には多様な表記バリエーションに対応できないという問題がある．たとえば，「もしもし」という語に対して「もしも～し」「もしも～～～し」などといった表記を仮に獲得することができたとしても，「もしも～～し」や「も～しも～し」という表記を獲得していなければ，これらの表記に対応することはできない．未知語モデルを学習する手法は，このような表記バリエーションに対応できる可能性はあるが，大量の学習データが必要となり，また，解析速度も一般的に大きく低下する．

さらに，いずれの手法にも元となった語に関する知識を活用できないという問題がある．人間が「ぉぃしぃ」や「もしも～～～し」などといった文字列をどのように理解するかを考えると，仮にそれらの文字列を見るのが初めてであったとしても，それらを完全に未知の文字列として語の区切りや意味を推定するのではなく，それぞれ「おいしい」，「もしもし」から変形した語として語を区切り，意味を理解していると考えられ，計算機による解析を行う場合も既知語に関連付けて解析するのが自然であると考えられる．

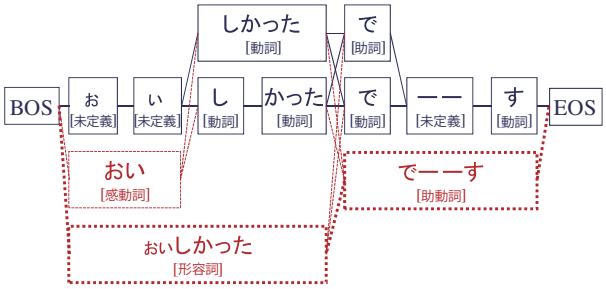

図 -3　形態素解析における崩れた表記の認識

■形態素解析における崩れた表記の認識

そこで JUMAN7.0 では，一般の未知語処理☆4 とは別に，形態素解析時に崩れた表記を既知語と関連付けることで，崩れた表記に対処している．具体的には，形態素解析において各位置から始まる形態素の辞書引きを行う際（図 -1 における手順①），入力文字列に対する辞書引きに加えて，表 -2 に挙げたような正規化ルールに基づいて入力文字列を変形し，変形した文字列に対しても辞書引きを行うことで，崩れた表記に対応している．

たとえば，先述の「ぉぃしかったでーーす」というテキストの解析を行う場合，長音記号を削除した文字列「ぉぃしかったです」や，小書き文字を置換した文字列「おいしかったでーーす」に対しても辞書引きを行い，図 -3 において破線で示された「ぉぃ」（感動詞），「ぉぃしかった」（形容詞），「でーーす」（助動詞）などの形態素候補を形態素ラティスに追加し，最終的に「ぉぃしかった」（形容詞）と「でーーす」（助動詞）の 2 形態素に正しく分割できるようになっている．JUMAN7.0 で採用されている手法の主な特長として次の 4 点がある．

1. 任意の文字数の挿入に対処できる
2. 学習データを必要としない
3. 解析速度にほとんど影響がない
4. 既知語に関する知識を活用できる

まず，入力文字列を規則的に変形した上で辞書引きを行っているため，多くの表記バリエーションに

☆4 JUMAN7.0 では「ググる」のような一般の新語・専門用語等への対応は人手で行うのではなく自動獲得によって行うべきであるとの考えに基づき，コーパスから自動的に構築した辞書を付属している．詳細は文献 3）を参照のこと．

タイプ	1万文あたりの改善数	1万文あたりの悪化数	解析速度の低下率	解析が改善する入力の例
長音記号の挿入	108	0	2.0%	・もしも〜し ・ぜーんぶ
小書き文字の挿入	36	0	0.4%	・行くぞぉ ・コチラでぇす
長音記号による置換	51	1	0.8%	・うらやまし〜 ・いー感じ
小書き文字による置換	137	2	0.7%	・ばぁちゃん ・書いていぃ

表 -3　崩れた表記の解析精度と速度低下率，および，解析が改善する入力の例

対応できる．たとえば，「もしも〜し」や「もしも〜〜〜し」，「も〜しも〜し」などといった表記があったとしても，いずれも「もしもし」という文字列に直した上で辞書引きを行うため，いずれの表記に対しても同じように対処することが可能である．

また，基本的に辞書引きの方法を改良しているだけであるため，学習データを必要としない．JUMAN7.0 では，通常の辞書引きにより生成された形態素候補を，正規化ルールを適用することで新たに追加された形態素候補より優先するため，後者に一定のペナルティコストを与えているが，このコストの設定に必要となる事例はごく少数である．

さらに，正規化ルールが適用された場合のみ新たな辞書引きを行うようにすることにより，解析速度の低下を最小限に抑えることができる．実際に Web テキストを解析した場合の解析速度の低下率を，1 万文あたりの改善数，悪化数，改善例とともに表 -3 に示す．

JUMAN7.0 で採用されている手法には，崩れた表記を辞書に載っている一般的な表記と関連付けることから，既知語に関する既存の知識を活用できるという特長もある．すなわち，崩れた表記で出現した場合であっても，JUMAN が使用している辞書に付与されているカテゴリやドメイン，反義語などの情報にとどまらず，大規模なシソーラスなど元となった語に関する種々の言語リソースを活用することが可能となる．

本章で紹介した手法は，文字列の削除・置換を行うためのコストを設定する必要はあるものの，基本的に辞書引きを改善することにより崩れた表記への対処を行っている．このため，各形態素の生起コストや連接コストの設定・推定法とは独立しており，ChaSen や MeCab などのような機械学習により各種コストを推定するシステムにも応用することが可能であることを最後に付け加えておく．

応用事例：言語学研究の支援

ここまで，辞書に含まれていない表現の解析に関する研究紹介を行ってきた．それでは，これらの表現を正しく解析できるようになったとして，その先にはどのような応用可能性が考えられるのであろうか．ここでは，新語処理技術の応用事例として，言語学研究支援に関する試みを紹介する [4]．

言語学とは，言うまでもなく，種々の言語現象の理論化を目的とした学問である．言語現象の分析を行う上で，大きな問題となるのが，言語データの包括的な収集が困難なことである．たとえば，言語学者が新語に関する研究を行おうとした場合，世の中で使われている新語に関するデータ収集を行う必要が生じる．しかしながら，これが非常に困難な作業であることは論を俟たない．

こうした問題解決のため，時系列 Web テキストと新語処理の技術を利用することにより，新語の言語学的分析の支援を行う研究を進めている．Web テキストには新語の用例が豊富に存在する．そうしたテキストを言語処理技術を用いて解析し，新語の用例を大規模に収集することによって，分析作業の網羅性および効率性を向上させることが狙いである．

言語学研究支援の一環として，新動詞の通時変遷を分析するための基盤構築に取り組んできた．図 -4 は，新動詞「ファブる」と「バルビる」の使用頻度の時間変化を比較したものである☆5．この図から，新動詞「ファブる」は世の中に広まりつつ

☆5　それぞれ「ファブリーズ（消臭剤の名称）を使う」「バルビレッジ（ゲームの名称）をプレイする」という意味の新語である．

あるが，逆に「バルビる」の流行は一過性であったことを見てとることができる．

このように大規模な Web テキストと新語処理技術を組み合わせることによって，新語の盛衰という，これまで観測困難であった言語データを簡単に取得できるようになることがお分かりいただけたかと思う．これは新語処理技術の応用の一例であるが，CGM を中心とする新世代のテキストメディアと言語処理技術の融合によって，新たな価値や知識が創出される可能性を感じ取っていただければと思う．

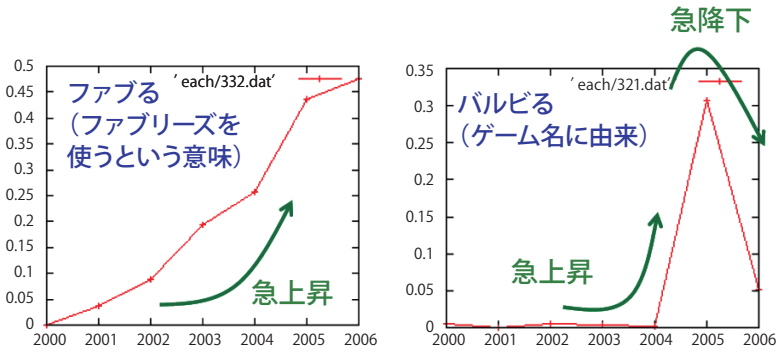

図 -4　新動詞「ファブる」と「バルビる」の使用頻度の通時変化の比較．縦軸は使用頻度，横軸は時間を表す．

実世界の言語処理に向けて

本稿では，新しい語や崩れた表記の扱いに焦点を当てて，従来技術では対処しきれなかった言語表現の形態素解析処理に関する取り組み，およびその応用事例を紹介した．

こうした表現は，従来の自然言語処理研究においては例外的な言語現象とみなされ，議論の中心となることは少なかった．しかし，これは強い言い方をするならば，新語や崩れた表記が存在しない，現実世界から乖離した言語が，自然言語処理の対象として暗黙のうちに想定されてきたということでもある．ブログを始めとする CGM の出現は，こうした問題について我々研究者が再考する良い機会であると言える．

CGM テキストは，少なくとも当面の間は自然言語処理において重要な研究対象であり続ける可能性が高い．そのため，現在の自然言語処理技術が処理対象として想定している言語と，現実世界において使用されている言語の差異を埋めていくことが，今後ますます重要になってくると考えられる．新語や崩れた表記に関する一連の研究が，そのような潮流を形成する一助となればと思う．

参考文献

1）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).
2）黒橋禎夫，河原大輔：日本語形態素解析システム JUMAN version 7.0 使用説明書，京都大学大学院情報学研究科 (2012).
3）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).
4）宇野良子，鍛治伸裕，喜連川優：新動詞の成立にみる意味と形の変化の相関―「ファブる」と「モフる」の分析から―，日本認知言語学会論文集第 10 巻 (2010).

（2011 年 11 月 18 日受付）

笹野 遼平（正会員）sasano@pi.titech.ac.jp

2009 年東京大学大学院情報理工学系研究科博士課程修了．博士（情報理工学）．2010 年より東京工業大学精密工学研究所助教．自然言語処理，特に述語項構造解析，照応解析の研究に従事．

鍛治 伸裕（正会員）kaji@tkl.iis.u-tokyo.ac.jp

2005 年東京大学大学院情報理工学系研究科博士課程修了．情報理工学博士．現在，東京大学生産技術研究所特任助教．自然言語処理の研究に従事．