論　文

大規模ウェブテキストからの片仮名用言の自動獲得

鍛治　伸裕[†a)]　　福島　健一[††b)]　　喜連川　優[†c)]

Acquisition of Katakana Verbs and Adjectives from Large Web Text

Nobuhiro KAJI[†a)], Ken'ichi FUKUSHIMA[††b)], and Masaru KITSUREGAWA[†c)]

あらまし　テキストマイニングでは，自然言語処理分野の基礎技術である形態素解析がモジュールとして利用されることが多い．しかし，ウェブには口語体のテキストが多く，新聞記事のような整ったテキストを対象としてきた自然言語処理技術では，十分な精度で解析を行うことは難しい．本論文では，形態素解析の精度低下は「ググる」などの片仮名用言が一因となっていることに着目し，それを大規模なウェブテキストから自動獲得する手法を提案する．

キーワード　ウェブ，自然言語処理，形態素解析

1. はじめに

近年，ブログなどの CGM (Consumer Generated Media) の爆発的な普及に伴って，評判情報の抽出など大規模ウェブテキストからの価値創出はますます重要な研究課題となっている [20], [24]．ウェブテキストのマイニングは，CGM が消費者の生の声と考えられることから，特にマーケティングにおいて大きなニーズが存在する．その他にも，企業や政府などの組織における意思決定支援に対するニーズも大きい．

ウェブテキストマイニングでは，自然言語処理分野の基礎技術である形態素解析 [3], [19] がモジュールとして利用されることが多い．しかし，ウェブには口語体のテキストが多く，新聞記事のような整ったテキストを対象としてきた自然言語処理技術では，十分な精度で解析を行うことは難しい．例えば，形態素解析器を用いてテキストからキーワードを抽出する場合には，白井らが指摘するように形態素解析器のエラーを無視できない [23]．また，乾らは，ウェブテキストを評判分析に活用することを念頭においたうえで，これを頑健に解析する技術の必要性を論じている [16]．

† 東京大学生産技術研究所，東京都
IIS, University of Tokyo, Tokyo, 153-8505, Japan
†† グーグル株式会社，東京都
Google, Inc., Tokyo, 150-8512, Japan
a) E-mail: kaji@tkl.iis.u-tokyo.ac.jp
b) E-mail: keni@google.com
c) E-mail: kitsure@tkl.iis.u-tokyo.ac.jp

ウェブテキストの解析を困難にしている言語現象の一つが新造語である．ウェブテキストでは「ググる」のような新しい言葉が日々生産されるため，新造語を頑健に解析することが必要となる．しかし，新造語は「サボる」のように日本語に定着してしまったもの(注1)を除けば，形態素解析辞書にほとんど登録されておらず，既存の形態素解析器では適切に処理できない．

上記の問題は，基本的には解析辞書を拡張することによって解決することができる．しかし，人手で拡張するという方法では作業コストが問題となる．新造語は日々作られては消えていくものなので，それに追従して辞書を整備するコストは大きい．さらに，特殊な単語はそもそも辞書作成者の意識に上らないということも，人手で辞書拡張を行うさいの問題である．例えば，技術者でなければ「サチる」という動詞の存在を知らないであろう．その他にも，一部のコミュニティでは「タヒる」という動詞が使われているが，それを知っている人間は多くないと思われる．

本論文では，新造語の中でも「ググる」などの片仮名用言に着目して，それを自動獲得する手法を提案する．新造語の獲得は，従来の自然言語処理技術では実現困難であったが，大規模なウェブテキストを活用することによってこれが可能になることを示す．提案手法によって，片仮名用言を解析辞書に自動的に登録す

(注1)：「サボる」は怠業を意味するフランス語の sabotage に由来する造語とされる．

ることが可能となり，くだけたウェブテキストの解析に対する頑健性が向上する．このことは，意見や評判の抽出といったアプリケーションにおいて特に重要である．例えば物の評判やその根拠は，名詞単独で表現されることはまれであり，名詞，動詞，形容詞が組合わさったフレーズで記述されることが多い．したがって，片仮名用言の自動獲得はこのようなアプリケーションの品質向上に寄与すると考えられる．

本稿の構成は以下のようになっている．まず 2. 節では片仮名用言の具体例をいくつか紹介し，その特徴について議論を行う．そして 3. 節では，2. 節の議論にもとづいて，片仮名用言を自動獲得する手法を説明する．4. 節では実験結果について述べる．獲得された片仮名用言の精度を調査し，さらに既存の形態素辞書との比較を行ったのでその結果を報告する．5. 節と 6. 節では，今後の課題と関連研究について論じる．最後に 7. 節でまとめを行う．

2. 片仮名用言

本論文では，片仮名用言を「語幹が片仮名で活用語尾が平仮名である用言 (動詞または形容詞)」と定義する．以下に片仮名用言の具体例を示す．なお，論文中で例として使用されている片仮名用言の意味は付録に記載されているので，不明なものは随時参照されたい．

(1) a. 鯉のエサについて ググって みる．
    b. 友人と一緒に デニる．
    c. モフり がいのありそうなおなか．
    d. 今日は徹夜で ハラシマる ぞ．
(2) a. プログラムが ヘサい ので公開したくない．
    b. ここの式変形が相当 テクい．
    c. 飯食うのも マンドい．

片仮名用言を自動獲得するにあたって，以下に述べるような 2 つの特徴に着目した．まず，片仮名用言は語幹が片仮名であるため，単語境界の判定が容易である．これは「ググる」と平仮名表記の「ぐぐる」を比較すると分かりやすい．

(3) a. 鯉のエサについてググってみる．
    b. 鯉のエサについてぐぐってみる．

日本語では，単語境界と文字種の切れ目が一致することが多い．そのため，(3a) では「ググって」の直前に単語境界があることが分かるが，(3b) ではそれは必ずしも自明でない．

次に，片仮名用言は語尾が規則的に変化するという特徴を持つ．そのため，所与の片仮名列に後続する文字列パターンを観測すれば，その片仮名列が用言の語幹であるかどうかを判定することができる．例として，動詞の語幹である片仮名列「ググ」に後続する文字列を以下に示す．

(4) a. 鯉のエサについて ググ ってみる．
    b. ググ らずわかったらマジすごいです．
    c. 分からないことはまず ググ れ．

片仮名列「ググ」の直後には「って」「ら」「れ」などの動詞語尾，及びそれに付属する助動詞「ず」などが出現している．このような出現パターンは動詞に特徴的であるため「ググ」は動詞の語幹であると推定できる．形容詞の場合も同様である．比較のため，名詞である片仮名列「グーグル」に後続する文字列の例を以下に示す．

(5) a. 検索エンジンはいつも グーグル などを使っています．
    b. グーグル の T シャツを愛用する．
    c. 今日から グーグル で働きます．

名詞である「グーグル」には，助詞とその組み合わせである「などを」「の」「で」などが後続しており，動詞の語幹である「ググ」とは異なるパターンを呈していることが確認できる．なお，動詞は活用型によって語尾の変化の仕方が異なるが，新造語の活用型はラ行五段活用が多いため，以下の議論ではラ行五段活用だけを考慮する．

3. 片仮名用言の自動獲得

次に，提案する片仮名用言の自動獲得手法を説明する．獲得の流れは以下の通りである．

（1） ウェブテキストから全ての片仮名列を抽出する．前節で述べたように，日本語では単語境界と字種の切れ目が一致しやすいので，抽出した片仮名列は全て片仮名用言の語幹候補と考える．ウェブテキストは田村らが構築したウェブアーカイブのデータを利用した [21]．

（2） 所与の片仮名列が動詞の語幹であるか否かを判定する分類器と，形容詞の語幹であるか否かを判定

する分類器を学習する．分類器には Support Vector Machine (SVM)，素性には片仮名列に後続する文字列を用いた．

（3） 抽出した片仮名列に対して上記の 2 つの分類器を順に適用し，動詞または形容詞の語幹と判定されたものから片仮名用言を獲得する．このとき，1 つの片仮名列から動詞と形容詞が同時に獲得されることもありうる．例えば片仮名列「テク」からは，動詞「テクる」と形容詞「テクい」の 2 つが獲得される．

（4） ここまでの処理で獲得された片仮名用言には，後述するようにノイズが混入しているため，後処理でこれを取り除く．

以下では SVM の学習に必要な訓練事例の構築方法，素性設計について述べる．そして最後に後処理を説明する．

3.1 訓練事例の自動構築

SVM の学習を行うためには大量の訓練事例が必要となる．これを人手で作成するコストを削減するために，本論文では，訓練事例をウェブテキストから自動構築する方法を提案する．以下では，片仮名列が動詞の語幹であるか否かを判定する SVM を学習する場合を考え，それに必要な訓練事例を自動構築する方法を説明する．ただし，形容詞の場合も動詞と全く同様の方法をとることができる．

我々は，既存の形態素解析器は解析辞書に登録されている単語であれば，90%後半という非常に高い精度で解析できることに着目した [3]．そこで，ウェブテキストを既存の形態素解析器(注2)で処理して，その結果から辞書登録語と解析された部分だけを取り出し，それを擬似的な訓練事例として使うことにした．

訓練事例の構築手順は次のようになる．まず，ウェブテキストを解析した結果から，辞書に登録されているラ行五段活用動詞と解析された部分を抽出する．そして，各動詞の語幹を擬似的な正例とする．素性設計についてはこの後で詳しく述べる．負例は，ラ行五段活用以外の動詞，名詞，形容詞，形容動詞，副詞から同様に作成する．

3.2 素 性 設 計

片仮名用言の語幹候補である片仮名列，および訓練事例 (前述の動詞語幹など) は素性ベクトルとして表現される．以下では前者を素性ベクトルに変換する方法を述べる．後者については同様なので省略する．

すでに述べたように，片仮名用言は語幹に後続する文字列に特徴を持つ．そこで，片仮名文字列に後続する平仮名 n-gram を 2 値素性として用いた．すなわち，ウェブテキストにおいて，所与の片仮名文字列にその平仮名 n-gram が一度でも後続して出現していれば 1，それ以外は 0 とする．なお，n-gram は 1-gram から 5-gram までを用いた．

素性は 2 値ではなく実数値にすることも考えられるが，予備実験では 2 値素性のほうが精度が良かったため，ここでは 2 値素性を採用した．また，素性を文字 n-gram ではなく平仮名 n-gram としているのは，素性数を減らして計算効率を良くするためである．この素性で捉えたいのは，主に動詞語尾の活用とそれに付属する助動詞などの情報なので，平仮名に限定しても十分有効に働くことが期待できる．

片仮名用言の語幹候補である片仮名列は品詞が曖昧な場合がある．例えば片仮名列「メタボ」を処理することを考える．これは動詞「メタボる」の語幹だけでなく，そのまま名詞として使うこともできる．そのため，もし「メタボ」が名詞として使われている回数が多ければ，SVM が「メタボ」を動詞の語幹と判定できなくなる可能性がある．そこで，全ての n-gram を素性として用いるのではなく，訓練事例の正例側に出現した n-gram のみを使うという素性選択を行った．これによって，ウェブテキスト中で「メタボ」が名詞として使われていたとしても，それは素性には反映されなくなる．

3.3 後 処 理

予備実験の結果，単純に SVM を適用するだけでは，獲得された片仮名用言にノイズが混入することが分かった．その原因は，片仮名用言の直前に片仮名表記の名詞 (または副詞) が出現していた場合に，それらを 1 つの用言として獲得していることであった．例えば，予備実験では「メチャウマい」が誤って形容詞として獲得されていた．しかし，これは形容詞ではなく，副詞「メチャ」と形容詞「ウマい」から構成される形容詞句である．

そこで，ノイズとなっているものは「名詞/副詞＋用言」という形をしていることを利用して後処理を行う．もし「ウマい」と「メチャウマい」のように，一方が他方を包含しているような 2 つの片仮名用言が獲得された場合，包含している方 (この場合は「メチャウマい」) は「名詞/副詞＋形容詞」という句である可能性が高いため削除する．

(注2)：http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

4. 実験と議論

4.1 デ ー タ

実験で用いた片仮名文字列の数は異なりで2,576,130であった．また，自動作成した訓練事例の総数は103,269であった．表1に訓練事例における品詞の内訳を示す．動詞の語幹を判定するSVMを学習するときには，動詞(ラ行五段)の2,527事例が正例として使われ，残りが全て負例として使われる．一方，形容詞の語幹を判定するSVMを学習するときには，形容詞の1,497事例が正例で，残りが全て負例となる．素性数は，素性選択を行わない場合で約380万，素性選択後で約30万となっている．

表1 訓練事例数
Table 1 Number of the training examples.

品詞	事例数
名詞	75,909
動詞 (ラ行五段)	2,527
動詞 (上記以外)	8,732
形容詞	1,497
形容動詞	3,760
副詞	10,844

4.2 獲 得 結 果

実験で獲得された用言の数を表2に示す．素性選択と後処理の効果を調べるために，それらを行わなかった場合の結果も併せて記載する．なおSVMの実装はSVM lightを用いた(注3)．全ての実験でカーネルは線形カーネルを使った．

表2 片仮名用言の獲得数
Table 2 Number of the acquired katakana verbs and adjectives.

素性選択	後処理	動詞	形容詞
○	○	880	210
○	×	1,021	220
×	○	297	116
×	×	301	116

次に，獲得された用言の精度を表3に示す．精度は，システムが出力した全ての用言を人手で調査して求めた．なお，再現率による評価は，正確な値を求めることは困難であるため，ここでは行わない．
素性選択と後処理によって明らかな精度向上が見られており，いずれも有効に働いていることが確認できる．素性選択を行わずに後処理のみ行った場合では，精度向上はわずかなものになっているが，これは獲得された片仮名用言の数が少なく，後処理が十分に機能しなかったためであると考えられる．

表3 片仮名用言の獲得精度
Table 3 The accuracy.

素性選択	後処理	動詞	形容詞
○	○	.768	.600
○	×	.684	.581
×	○	.475	.543
×	×	.468	.543

獲得例を調べると,「ミネラる」のように，本来なら片仮名で表記ですべき「ル」が平仮名で誤表記されている名詞が，誤って動詞として獲得されている例がいくつか見られた．こうした失敗例をなくすためには，ノイズの多いウェブテキストを何らかの方法でクリーニングするが必要となると考えられる．一方，形容詞の失敗例には，例えば「イケテナい」や「ワカラナい」のような「全て片仮名表記された動詞＋形容詞と同じ活用をする接尾辞」が見られた．これは後続文字列を利用するだけでは対処することが難しいため，新しい素性の導入などが必要であると考えられる．

提案手法のように機械学習を使うのではなく，単純な規則を用いた手法との比較を行った．ウェブテキストから「片仮名列＋る」「片仮名列＋い」というパターンにマッチする文字列を，それぞれ動詞，形容詞として抽出して，その精度を求めた．ただし，パターンにマッチした全ての文字列を人手で調べることは困難なので，それぞれ100個ずつを無作為に選択して評価を行った．その結果，動詞は100個中53個，形容詞は6個が適切なものであった．さらに，片仮名動詞を既存の形態素解析器で解析すると「未知語＋る (助動詞)」と解析されることから，これをルールとして抽出するという方法を試した．片仮名形容詞に対しても，同様に「未知語＋い (動詞)」という解析結果を抽出するという方法を試した．それぞれの結果から100個ずつを無作為に選択して評価を行ったところ，動詞は100個中2個が適切なものであり，形容詞は全て誤りであった．これらの結果から，片仮名用言の抽出は，単純な規則を適用するだけでは十分な精度を達成できないことが分かった．

4.3 既存の形態素解析辞書との比較

片仮名用言を自動獲得する最終的な目的は，形態素解析辞書の拡張を行うことであった．そのため，自動獲得の精度だけでなく，獲得できた辞書未登録語の数

(注3)：http://svmlight.joachims.org

による評価も重要である．

そこで，自動獲得した片仮名用言と既存の形態素解析辞書の比較を行い，どの程度，辞書未登録語を獲得できたかを調べた．比較には，前述の実験で適合率を調査した過程において，正しい片仮名用言であると確認されたものだけを用いた．その数は動詞が 736，形容詞が 148 である．また，形態素解析辞書は JUMAN(version 6.0) と IPADIC(version 2.7) を用いた．

表 4 に，各辞書に未登録であった片仮名用言の数を示す．この表から，自動獲得した動詞の 80%以上と形容詞の 50%以上が，既存の形態素解析辞書に未登録であったことが確認できる．したがって，本手法は形態素解析辞書の拡張に有効であると結論づけることができる．

表 4　形態素解析辞書に未登録であった片仮名用言の数
Table 4　Number of katakana verbs and adjectives that are not in the dictionaries.

	動詞	形容詞
JUMAN	638	94
IPADIC	622	84

以下に，IPADIC に未登録であった片仮名用言の一部を示す．辞書を自動拡張することによって，このような用言が正しく解析できるようになる．

- 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
- オイチい，スンバラシい，ウッザい，ナヨい，ヘヴィい，ズブい

5. 今後の課題

実験の結果，片仮名用言の獲得精度は動詞で 76.8%，形容詞で 60.0%であり，獲得した用言をそのまま解析辞書に登録するには精度が不十分であることが予想される．そのため，今後は手法のさらなる改良を行い，精度を向上させる必要がある．改良方法としては，例えば SVM 以外の学習アルゴリズムの適用や，新しい素性の導入などが考えられる．また，提案手法は片仮名用言に特化したものとなっているため，名詞や副詞など他の品詞を扱うことができない．そのため，提案手法を他品詞に拡張することも重要な課題であると考えている．

獲得された片仮名用言の中には，ウェブテキストにおける用例を一読しただけでは，その意味をすぐに把握できないものも存在した．

(6) a. わざと イナタい メロディを吹いてみたりとか．
　　b. 心温まる イナタい サウンド．

日本語を母国語とする者であっても，上記の用例のみから「イナタい」の意味を推測することは簡単ではない．今後は，獲得した片仮名用言の説明文を自動生成する方法を検討したい．用語説明文の自動生成にはすでに研究が存在するので，これと同様の手法が適用できる可能性がある [17]．

6. 関連研究

自然言語処理の分野では，形態素解析辞書に登録されていない語は未知語と呼ばれており，形態素解析の精度を低下させる大きな要因とされてきた．これまでにも，形態素解析辞書の拡張を目的として，未知語の自動獲得を行った研究が多く存在する．

森らは文字 n-gram にもとづく未知語抽出手法を提案した [5], [18]．品詞タグ付きコーパスから名詞や動詞など各品詞の特徴ベクトルを求め，それとのユークリッド距離を測定することによって，ある文字列が未知語であるかどうかを算出している．特徴ベクトルには，単語の前後に出現する文字 n-gram の頻度が用いられている．本研究は，文字 n-gram を利用するという点で森らの研究と類似点が見られるが，片仮名用言の抽出に特化しているという点が異なる．また，提案手法は，人手でタグ付けされた訓練事例を必要としない点も大きな特徴となっている．

Asahara らは，未知語抽出を文字レベルの chunking 問題として定式化した [1]．そして，文字種，形態素解析器の n-best 解などを素性として利用して，SVM にもとづく chunker の学習を行った．

Nakazawa らは片仮名語には外来語が多いことに着目し，日本語コーパスだけでなく英和辞書，和英辞書，英語コーパスを併用することによって片仮名語の自動獲得を行った [9]．Nakazawa らの手法は，コーパス以外に辞書を利用している点において，本手法とは方向性が大きく異なる．また，日本語ではなくドイツ語を対象とした研究であるが，Koehn らも Nakazawa らと類似の手法を提案している [2]．

一方，事前に形態素解析辞書を拡張するのではなく，辞書に登録されていない単語でも適切に処理できるように，形態素解析モデルを改良する試みも存在する．

これは辞書の拡張を行う提案手法と全く異なるアプローチであるが，両手法は容易に統合可能であるため非常に興味深い．

Nagata は，確率的生成モデルの枠組みを用いて，未知語処理を行うことができる形態素解析モデルを提案した [6]．未知語の生成確率を，その単語を構成する文字列が生成される確率として定義することによって，未知語処理を実現している．

Uchimoto らは，Maximum Entropy Markov Model (MEMM) にもとづく形態素解析モデルを提案した [11]〜[13]．識別モデルの一種である MEMM を用いることによって，文字種や語頭文字列や語末文字列など，従来の生成モデルでは利用困難であった重複素性を取り込むことによって，未知語処理の解析精度の向上を実現している．

Nakagawa は，2 つの形態素解析モデルを組み合わせたハイブリッド法を提案した [7]．既知語を高い精度で解析できるマルコフモデルと，未知語に対して頑健な文字ベースの Maximum Entropy (ME) モデルを組み合わせることによって，個々のモデルよりも高い解析精度を達成している．また，別の文献において Nakagawa らは，未知語の品詞推定を行う手法も提案している [8]．しかし，この方法は，テキスト中の未知語の箇所が同定済みであることなど，いくつかの強い仮定にもとづいており，ここまで議論してきた未知語処理とは問題の設定が異なっている．

東らは，Uchimoto らが未知語処理に MEMM を導入したのに対して，Conditional Random Field (CRF) の適用を提案した [22]．MEMM のように識別モデルを段階的に適用するモデルでは label bias と length bias [3] と呼ばれる問題が発生するという観察にもとづき，これらの問題に対して耐性を持つ CRF を適用することによって，未知語の解析精度を向上させることに成功している．

未知語問題は，日本語以外でも，単語境界が明示されていない言語であれば発生する．そのため，アジア言語，特に中国語においても未知語処理の研究は盛んに行われてきた．日本語と同様に，コーパスから未知語の獲得 [4], [14] と，未知語に対して柔軟に対応できる解析モデル [10], [15] の両方の側面から研究が進められている．

7. おわりに

ブログなど CGM の爆発的な普及に伴って，近年，大規模なウェブテキストの解析はますます重要性を増してきている．しかし，ウェブテキストのような口語体の文章は，従来の形態素解析技術では十分な精度で解析することはできない．この問題の解決を図るため，形態素解析辞書を拡張することを目的として，片仮名用言の自動獲得手法を提案した．実験の結果，本手法は従来の形態素解析辞書には登録されていなかった片仮名用言を多数獲得できることが分かった．これによって，形態素解析辞書を自動拡張する見通しを得た．

文　献

[1] Masayuki Asahara and Yuji Matsumoto. Japanese unknown word identification by character-based chunking. In *Proceedings of the International Conference on Computational Linguistics*, pp. 459–465, 2004.

[2] Philip Koehn and Kevin Knight. Empirical methods for compound splitting. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pp. 187–193, 2003.

[3] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, 2004.

[4] Hongqiao Li, Chang-Ning Huang, Jianfeng Gao, and Xiaozhong Fan. The use of SVM for Chinese new word identification. In *In Proceedings of the International Joint Conference on Natural Language Processing*, pp. 497–504, 2004.

[5] Shinsuke Mori and Makoto Nagao. Word extraction from corpora and its part-of-speech estimation using distributional analysis. In *Proceedings of the International Conference on Computational Linguistics*, pp. 1119–1122, 1996.

[6] Masaaki Nagata. A part of speech estimation method for Japanese unknown words using a statistical model of morphology and context. In *Proceedings of the Association for Computational Linguistics*, pp. 277–284, 1999.

[7] Tetsuji Nakagawa. Chinese and Japanese word segmentation using word-level and character-level information. In *Proceedings of the International Conference on Computational Linguistics*, pp. 705–712, 2004.

[8] Tetsuji Nakagawa and Yuji Matsumoto. Guessing parts-of-speech of unknown words using global information. In *Proceedings of the International Conference on Computational Linguistics*, pp. 705–712, 2006.

[9] Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. Automatic acquisition of basic katakana lexicon from a given corpus. In *Proceedings of the International Joint Conference on Natural Language*

Processing, pp. 682–693, 2005.

[10] Fuchun Peng, Fangfang Feng, and Andrew McCallum. Chinese segmentation and new word detection using conditional rnadom fields. In *In Proceedings of the International Conference on Computational Linguistics*, pp. 562–568, 2004.

[11] Kiyotaka Uchimoto, Chikashi Nobata, Atsushi Yamada, Satoshi Sekine, and Hitoshi Isahara. Morphological analysis of the spontaneous speech corpus. In *Proceedings of the International Conference on Computational Linguistics*, pp. 1298–1302, 2002.

[12] Kiyotaka Uchimoto, Chikashi Nobata, Atsushi Yamada, Satoshi Sekine, and Hitoshi Isahara. Morphological analysis of a large spontaneous speech corpus in Japanese. In *Proceedings of the Association for Computational Linguistics*, pp. 479–488, 2003.

[13] Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. The unknown word problem: A morphological analysis of Japanese using maximum entropy aided by a dictionary. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 91–99, 2001.

[14] Andi Wu and Zixin Jiang. Statistically-enhanced new word identification in a rule-based Chinese system. In *Proceedings of SINGHAN*, pp. 46–51, 2000.

[15] Yue Zhang and Stephen Clark. Joint word segmentation and pos tagging uisng a single perceptron. In *In Proceedings of the Association for Computational Linguistics*, pp. 888–896, 2008.

[16] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, pp. 201–242, 2006.

[17] 桜井裕, 佐藤理史. ワールドワイドウェブを利用した用語説明の自動生成. 情報処理学会論文誌, Vol. 43, No. 5, pp. 1470–1480, 2002.

[18] 森信介, 長尾眞. n グラム統計によるコーパスからの未知語抽出. 情報処理学会論文誌, Vol. 39, No. 7, pp. 2093–2100, 1998.

[19] 浅原正幸, 松本裕治. 形態素解析のための拡張統計モデル. 情報処理学会論文誌, Vol. 43, No. 3, pp. 685–695, 2002.

[20] 倉島健, 藤村考, 奥田英範. 大規模テキストからの経験マイニング. 電子情報通信学会 第 19 回データ工学ワークショップ, 2008.

[21] 田村孝之, 喜連川優. 大規模ウェブアーカイブのための更新クローラの設計と実装. 日本データベース学会レターズ, Vol. 6, No. 1, pp. 173–176, 2007.

[22] 東藍, 浅原正幸, 松本裕治. 条件付確率場による日本語未知語処理. 情報処理学会研究報告, 自然言語処理研究会 (2006-NL-173), pp. 67–74, 2006.

[23] 白井智, 鳥井修, 金井達徳. 反復文字列階層グラフによる文書からのキーワード自動抽出. 電子情報通信学会 第 16 回データ工学ワークショップ, 2005.

[24] 福原知宏, 中川裕志, 西田豊明. 感情表現と用語のクラスタリングを用いた時系列テキスト集合からの話題検出. 人工知能学会全国大会, 2006.

付　　録

表 A·1　本文中で使われた片仮名用言の意味 (五十音順).
Table A·1　Meanings of katakana verbs and adjectives used in the paper.

ぐぐる 【ググる】 (動詞・ラ行五段活用) （1）ウェブ検索エンジン Google を使うこと．まれに「Yahoo!でググれ」など，本来の意味から逸脱して，単に検索エンジンを使うという意味でも使われる．
たひる 【タヒる】 (動詞・ラ行五段活用) （1）頻脈になるという意味の隠語．独語の tachycardia に由来する．主に医師や看護師の間で使われる． （2）死ぬこと．漢字の「死」が片仮名の「タヒ」に見えることによる．
てくい 【テクい】 (形容詞) （1）テクニックを巧みに駆使して物事を行う様子．
てくる 【テクる】 (動詞・ラ行五段活用) （1）何かに対して高度なテクニックを有している者がそれを行使すること．または，その様子．
でにる 【デニる】 (動詞・ラ行五段活用) （1）ファミリーレストランチェーンのデニーズで食事をすること．
はらしまる 【ハラシマる】 (動詞・ラ行五段活用) （1）主に同人誌のコミュニティにおいて，原稿を書くこと．原稿の誤字である原縞が「はらしま」と読めることに由来する．
へさい 【ヘサい】 (形容詞) （1）物事が下手で格好悪いようす．
まんどい 【マンドい】 (形容詞) （1）面倒くさいの意.「面倒くさい → メンドい → マンドい」のように訛ったもの．(参考) マンドクサい．
めたぼる 【メタボる】 (動詞・ラ行五段活用) （1）メタボリック症候群 (metabolic syndrome) になること．また，そこから転じて太ること．
もふる 【モフる】 (動詞・ラ行五段活用) （1）仰向けになった猫の腹を撫でたり，腹の毛に顔を埋めて愛でること． （2）猫の冬毛が伸びてフサフサしてくること．

（平成 xx 年 xx 月 xx 日受付）

鍛治　伸裕

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．

福島　健一

2008 年東京大学大学院情報理工学系研究科修士課程修了．同年 4 月よりグーグル株式会社ソフトウェアエンジニア．自然言語処理，知識発見，機械学習等に興味を持つ．

喜連川　優

1978 東京大学工学部電子工学科卒業．1983 同大大学院工系研究科情報工学専攻博士課程修了．工学博士．同年同大生産技術研究所講師．現在，同教授．2003 より同所戦略情報融合国際研究センター長．データベース工学，並列処理，Web マイニングに関する研究に従事．現在，日本データベース学会理事，情報処理学会，電子情報通信学会各フェロー．ACM SIGMOD Japan Chapter Chair，電子情報通信学会データ工学研究専門委員会委員長歴任．VLDB Trustee(1997-2002), IEEE ICDE, PAKDD,WAIM などステアリング委員．IEEE データ工学国際会議 Program Co-chair(99),General Co-chair(05).

Abstract In Japanese text mining, word segmentation and POS tagging are indispensable. However, there are a lot of colloquial expressions in Web text and it is hard for conventional NLP techniques to process such text. Considering that the major source of the error arises from katakana verbs and adjectives, this paper presents a method of acquiring katakana verbs and adjectives from large Web text.

Key words The Web, NLP, Joint Word Segmentation and POS Tagging