情報処理学会論文誌　データベース　Vol. 3　No. 3　68–77 (Sep. 2010)

くだけた表現を高精度に解析するための正規化ルール自動生成手法

池田和史[†1]　柳原正[†1]
松本一則[†1]　滝嶋康弘[†1]

ブログ上の文書には口語的な表現や特有の表記などのくだけた表現が多数含まれるため，一般の形態素解析器を用いても十分な解析精度を得ることはできない．くだけた表現は人手により辞書登録されることが一般的であるが，人的コストの大きさや専門的な知識を必要とすることが課題である．本稿ではくだけた表現を正規な表現に修正することで高精度な形態素解析を実現する手法を提案する．提案手法ではくだけた表現の修正候補文字列をくだけた表現の少ない文書から自動的に検索し，修正ルールを生成する．生成した多数の修正ルールから文脈に適した修正ルールを選択的に適用するために，検索結果における修正候補文字列の出現頻度，修正前後の文字列間における編集距離，修正前後の文の形態素解析結果の比較，を用いて修正ルールをスコアリングする手法を合わせて提案する．提案手法と従来手法の性能比較評価実験を行い，各手法における未知語の出現率や単語区切りの正確さ，修正前後の文の意味変化を定量的に評価した．提案手法では従来手法と同程度の単語区切りの正確さを維持しながら，対象文章の未知語出現数を 36.1% 減少させることに成功した．これは従来手法における未知語減少数の 2.5 倍以上である．

Automatic Rule Generation Approach for Morphological Analysis of Peculiar Expressions on Blog Documents

Kazushi Ikeda,[†1] Tadashi Yanagihara,[†1]
Kazunori Matsumoto[†1] and Yasuhiro Takishima[†1]

In this paper, we propose an algorithm for reducing the number of unknown words on blog documents by replacing peculiar expressions with formal expressions. Japanese blog documents contain many peculiar expressions regarded as unknown sequences by morphological analyzers. Reducing these unknown sequences improves the accuracy of morphological analysis for blog documents. Manual registration of peculiar expressions to the morphological dictionaries is a conventional solution, which is costly and requires specialized knowledge. In our algorithm, substitution candidates of peculiar expressions are automatically retrieved from formally written documents such as newspapers and stored as substitution rules. For the correct replacement, a substitution rule is selected based on three criteria; its appearance frequency in retrieval process, the edit distance between substituted sequences and the original text, and the estimated accuracy improvements of word segmentation after the substitution. Experimental results show our algorithm reduces the number of unknown words by 36.1%, maintaining the same segmentation accuracy as the conventional methods, which is 2.5 times the reduction rate of the conventional methods.

1. まえがき

近年，インターネットの普及により，一般ユーザによる Web 上での情報発信の手段としてブログが注目されており，ブログを対象とした情報抽出や検索，ランキングなどに関する研究がさかんに行われている[1),2)]．しかし，ブログ文書には「うっそー」,「すごーい」のような口語的な表現や「ヵわいい」,「わたしは Uわ」（「わたしは」と読む）のような特有の表記などの **くだけた表現** が含まれ，その多くは一般の形態素解析器では未知の語として扱われるため，十分な言語解析精度を得ることができないという問題がある．現在ではくだけた表現を人手により辞書登録することが一般的であるが，未知語の登録には品詞や活用形の登録，既存の辞書との互換性の維持など，言語処理に関する専門的なスキルを必要とし，人的コストが大きい点が問題となる．著者らの経験では，1 人月あたり約 3 万種類の未知語登録が可能であるのに対し，ブログ 600 万文を著名な形態素解析器 MeCab[3)] を用いて解析したところ，約 65 万種類の未知語が検出されたことから，ブログ文書のくだけた表現を正しく解析することは困難といえる．

本稿におけるくだけた表現とは正規な表現から派生した表現と考えており，その派生の仕方にはいくつかの傾向が見られる．たとえば「ヵわいい」や「わたしは Uわ」のように形状が似ている文字の代替が起こりやすい傾向にある．他の傾向として,「うっそー」や「すごーい」のように会話における発音の変化傾向にあわせた表記がなされる．また,「かっこいい」がブログ上では「カッコイイ」と記載されるように，本来ひらがなで書かれるべき語を意図的にカタカナ表記にするなどの傾向があるが，文字形状の類似度を計るうえで，現状の OCR

†1 KDDI 研究所
KDDI R&D Laboratories, Inc.

（Optical Character Reader）などの認識精度は十分ではなく，口語表現を網羅するための辞書拡充に要する人的コストも大きい．

これらのくだけた表現の解析精度を向上させるため，本稿ではくだけた表現を正規な表現へと修正する手法を提案する．提案手法ではくだけた表現の修正候補文字列をくだけた表現の少ない新聞コーパスなどから自動的に検索し，文字列変換のルール（**修正ルール**）を生成する．たとえば，文字列「ヵわいい」を「かわいい」に変換することで，くだけた表現を正規な表現に修正できる．生成した多数の修正ルールから文脈に適した修正ルールを選択的に適用可能にするため，検索結果における修正候補文字列の出現頻度，修正前の文字列から修正後の文字列への文字列編集距離，修正前後の文の形態素解析結果の比較，という 3 つの指標を用いて修正ルールをスコアリングする手法を合わせて提案する．

提案手法を実装し，性能評価実験を実施した．修正ルールをスコアリングする 3 つの指標の重み付けパラメータと修正ルールを適用するスコアの閾値を調整することで，くだけた表現修正の再現率と適合率のトレードオフについて評価した．加えて，従来手法と提案手法の性能比較評価実験においては，形態素解析時の未知語の出現率や単語区切りの正確さ，修正前後における文の意味の変化について定量的に評価した．提案手法では，従来手法と同程度の単語区切りの正確さで従来手法の 2.5 倍以上の未知語を減少させることに成功した．これは評価対象文章の未知語出現数の 36.1% に相当する．

2. 関 連 研 究

チャットの口語的表現を対象とした形態素解析辞書拡張手法[4] では，チャットの文章を分析し，人手によって辞書拡張のルールを作成することで，既存の辞書から派生した語を辞書登録する．たとえば,「がっこう」は「がっこー」と表現されるなどの例から，直前の文字の母音が「o」の場合,「お，う，ー，〜」は互いに置換可能である，などのルールを提示している．この手法により，辞書登録の人的コストは軽減されるが，人手によるルール作成は作業者が参考にした文例に依存したり，主観に基づいたりしやすい．以前の著者らの研究報告[5] では文献 4) を参考にルールを作成し，ブログ 200 万文を形態素解析し，単語区切りに変化が見られた 53488 文のうち，600 文をサンプリングして評価したところ，37.2%の文はルール適用前と比べて単語区切りが悪化していることが確認された．

このほかにも口語的表現や話し言葉を言語的な観点などから分析した形態素解析精度向上のための手法が提案されている．文献 6) では,「〜しちゃう」などの口語特有の言い回しを分析し，人手により辞書登録を行うことで，口語の形態素解析精度が向上することが報告されている．同様に，文献 7)，8) では，話し言葉の形態素解析を対象とした研究成果が報告されており，様々な観点から口語的表現を分析し，特徴を列挙している．しかし，上記のような言語解析には専門的なスキルや多くの労力を要するなど，人的コストの大きさが課題となる．また，形態素解析における未知語の解消についてもさかんに研究が行われており，カタカナ語の表記の揺れを解消する手法[9] や Web から新語を獲得する手法[10]，未知語の品詞推定を行う手法[11]，単語分割境界を推定する手法[12] などが提案されている．これらの手法は未知語の解消に貢献するが，著者らが対象としているくだけた表現の修正を対象としたものではない．

これに対し，著者らが以前に提案した修正ルールの統計的スコアリグ手法[5] では，少数の汎用的な修正ルール（プリミティブルール）をあらかじめ人手により与える．プリミティブルールをもとに，特定の文脈でのみ利用できる特殊な修正ルールを生成し，大規模コーパスを用いて統計的にスコアリングすることで，文脈に応じた修正ルールの選択を可能にした．たとえば,「ゎ⇒は」と「ゎ⇒わ」という汎用的な修正ルールを人手により与えると，(a)「今日ゎ⇒今日は」や (b)「今日ゎ⇒今日わ」のような，より特殊な修正ルールを大規模コーパスから自動的に生成し，(a) の「今日ゎ⇒今日は」の方が統計的に正解率が高いことを学習する．しかし，この手法では与えられたプリミティブルールを組み合わせた修正ルールの生成に限定されるため,「困っちゃう」を「困ってしまう」に修正するためには「ちゃ⇒てしま」などの修正ルールを人手により与える必要があるなど，拡張性が低いことや修正結果が与えられたプリミティブルールに大きく依存してしまうという課題があった．

本稿で提案する手法は，(1) くだけた表現の修正候補文字列をくだけた表現の少ない文書から自動的に検索し，修正ルールを生成することにより，人手によるプリミティブルールの付与を必要としない点，(2) 検索結果における修正候補文字列の出現頻度と修正前後の文字列間における編集距離，修正前後の文の形態素解析結果の比較，の 3 つの指標を用いて修正ルールをスコアリングすることで，多数の修正ルールから文脈に適した修正ルールを選択的に適用可能な点，において文献 5) の手法とは大きく異なる．

3. くだけた表現の定義

本稿におけるくだけた表現の定義は，正規の表現が形態素解析の辞書に登録されているが，表記上の揺らぎなどによって未知語として扱われるもの，とする．このようなくだけた表現は正規の表現から多様な派生の仕方をするため，人手で個別に形態素解析の辞書に登録するよりも，提案手法のように正規の表現に戻して解析を行う方が適していると考えられる．

表 1　くだけた表現の分類と具体例
Table 1　Classification of peculiar expressions and their examples.

	大分類	小分類	具体例
くだけた表現	形状が似ている文字の代替	大文字を小文字で代替	ゎたしは（わたしは），かわぃぃヵも（かわいいかも），ょわった（よわった）
		記号による代替	わた∪は（わたしは），教えヶ下土い（教えて下さい），㊥ 学校（中学校），① 時に（1 時に）
		併せ字による代替	わナニしは（わたしは），ナょまえ（なまえ），└ノんご（りんご）
	発音の変化傾向にあわせた表記	文字の挿入	でっかい（でかい），はやーい（はやい），うそだあ（うそだ），えええええっ!?（えっ!?）
		文字の削除	どしたの？（どうしたの？），たいくの授業（たいいくの授業），おわた（おわった）
		文字の代替	けーたい（けいたい），すげぇ（すごい），かわいしゅぎる（かわいすぎる）
	ひらがな，カタカナ表記の代替		メッチャ（めっちゃ），カッコイイ（かっこいい），びたみん（ビタミン），
くだけた表現でない未知語	顔文字，アスキーアート		(^o^)，OTL，(・∀・) ニヤニヤ
	感情表現		（笑），(爆)，(泣，(汗
	擬音，擬態語		ぎとぎと，ぺっちょり，ぱかぱか
	固有名詞		りーしゃさんのブログは，ぐーたのエサ

くだけた表現は正規の表現からの派生の仕方で大きく 3 つに分類することができる．分類とそれぞれに分類されるくだけた表現の例を**表 1** に示す．具体的には，(1) 形状が似ている文字の代替が起こりやすい，(2) 会話における発音の変化傾向にあわせた表記がなされる，(3) ひらがな表記とカタカナ表記の代替が起こりやすい，といった傾向がある．(1) については，さらに 3 つの小分類があり，「わ」の代わりに「ゎ」，「か」の代わりに「ヵ」を利用するなど小文字による代替，「し」の代わりに「∪」，「中」の代わりに「㊥」を利用するなど記号による代替，「た」の変わりに「ナニ」の 2 文字を利用するなど併せ字による代替がある．(2) については，「でかい」は口語では「でっかい」と発音されることから，「っ」を挿入した表記とするものや「えええええっ!?」のように，同じ文字を重ねて臨場感を高める表現，「どしたの？」のような文字を脱落させた省略表現，「けいたい」を「けーたい」と表記するなどの代替表現がある．(3) については，「カッコイイ」のように，本来ひらがなで表記されるべき単語をカタカナで表記したり，反対に「びたみん」のように本来カタカナで表記されるべき単語をひらがなで表記するといった傾向がある．一方，本稿でくだけた表現として扱わない未知語として，顔文字やアスキーアート，感情表現，擬音，擬態語，固有名詞がある．これらについては正規の表現からの派生ではないため，個々に辞書登録を行うことを想定している．5 章における実験では，商用のブログ 5 万文に含まれる未知語数と，未知語中のくだけた表現数を計量する．

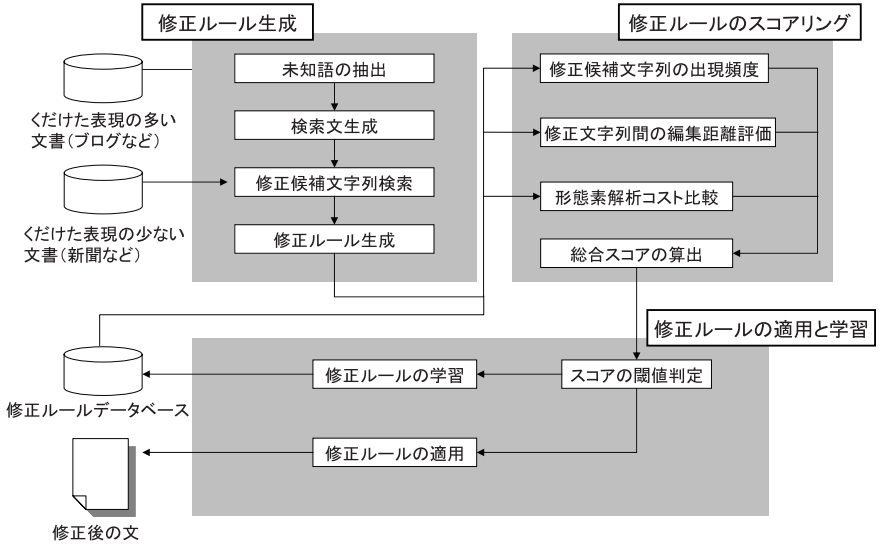

図 1　提案手法の全体像
Fig. 1　Overview of the substitution algorithm.

4. 提 案 手 法

提案手法の全体像を**図 1** に示す．提案手法では，くだけた表現を多く含むブログなどの文書を入力とし，くだけた表現の少ない新聞などの文書からくだけた表現の修正候補を自動的に検索し，修正ルールとして生成する．生成した修正ルールを 3 つの言語的な指標によりスコアリングすることで，得られた多数の修正ルールの中から文脈に適した修正ルールを選択することができる．以下では，各処理の詳細について説明する．

4.1　修正ルールの生成

修正ルールの生成はくだけた表現の抽出と修正候補文字列の取得により実現される．ここでは「できるかどう かわ 分かりません」というくだけた表現を含む文の修正を例にあげて

くだけた表現を含む文（修正対象となる文）：
「できるかどう ヵわ 分かりません」
形態素解析結果：
「できる/か/どう/ヵわ/分かり/ませ/ん」
未知語：ヵわ　…(1)
検索文生成：
「どう*分かり」　…(2)
くだけた表現の少ない文書からの検索結果：
「これかどう かは 分かりません」
「よくあるかどう か 分かりません」
「どう したらいいのか 分かりません」
「この先どう かは 分かりません」　…(3)
「本当かどう か 分かりませんが」
「使うかどう かは 分かりませんけどね」
「あるかどう かは 分かりません」
「どう なっているか 分かりませんよ」
修正ルールの生成：
「ヵわ⇒かは」
「ヵわ⇒か」
「ヵわ⇒したらいいのか」　…(4)
「ヵわ⇒なっているか」

図 2　修正ルール生成方法の具体例
Fig. 2　Generation of substitution rules.

説明する（**図 2**）．くだけた表現の多くは形態素解析辞書に登録されていないため，形態素解析時に未知語として検出される（図 2 (1)）．検出されたくだけた表現の修正候補文字列をくだけた表現の少ない新聞文書などから検索する．くだけた表現の未知語部分（図 2 では「ヵわ」）を任意の文字列（ワイルドカード）とし，未知語部分に隣接する文字列と合わせて検索文とする（図 2 (2)，ここでは隣接する 1 形態素ずつを検索文生成に利用した）．くだけた表現の少ない文書から修正候補文字列を検索し，取得する（図 2 (3)）．未知語部分から修正候補文字列への文字列変換を修正ルールとして生成する（図 2 (4)）．これにより，多数の修正ルールを自動生成できる．生成した多数の修正ルールのうち，文脈に適した修正ルールを選択的に適用するためのスコアリング手法について 4.2 節で説明する．

表 2　修正候補文字列の出現頻度に基づくスコアリング例
Table 2　Scoring based on appearance frequency.

修正ルール	出現頻度	出現頻度/検索件数
ヵわ ⇒ かは	4	0.5
ヵわ ⇒ か	2	0.25
ヵわ ⇒ したらいいのか	1	0.125
ヵわ ⇒ なっているか	1	0.125

4.2　修正ルールのスコアリング

提案手法では修正ルールのスコアリングに，(1) 検索結果における修正候補文字列の出現頻度，(2) 修正前後の文字列間における編集距離，(3) 修正前後の形態素解析コスト値の差分，の指標を用いる．以下では各スコアの算出方法について説明する．

4.2.1　修正候補文字列の出現頻度に基づくスコアリング

修正ルール生成時の検索結果における修正候補文字列の出現頻度を修正ルールのスコアリングに用いる．図 2 (3) における検索結果の出現頻度をまとめると**表 2** のようになる．出現頻度の高い文字列はくだけた表現が出現した文脈と類似した文脈でよく利用される表現であると考えられ，くだけた表現の修正候補文字列である可能性が高い．一方，類似した文脈であまり利用されていない表現は修正候補文字列である可能性が低い．評価値が検索件数に依存しないように，出現頻度を検索件数で割り，正規化して利用する．

4.2.2　修正前後の文字列間における編集距離に基づくスコアリング

くだけた表現は正規な表現から派生した表現であり，「すごーい」や「どぅかな」のように，正規な表現に対して少数の文字の挿入や削除，置換を行ったものであることが多い．ここで，文字列間の編集距離（レーベンシュタイン距離[13]）を考える．編集距離とは，2 つの文字列がどの程度異なっているかを表す指標であり，一方の文字列を他方の文字列に変換するために必要な挿入，削除，置換の最小コストとして与えられる．たとえば，挿入，削除のコストを 1，置換のコストを 2 とすると，「フォーラム」から「ファーム」への編集は「ォ」を「ァ」に置換し，「ラ」を削除する方法が編集距離 3 で最小となる．

編集距離を用いると，図 2 で生成した各修正ルールのスコアは**表 3** のようになり，「ヵわ⇒したらいいのか」や「ヵわ⇒なっているか」などは編集距離が大きいため，修正ルールのスコアは低くなる．また，くだけた表現では「ヤバい」や「カッコイイ」のように本来ひら

表 3　編集距離に基づくスコアリング例
Table 3　Scoring based on edit distance.

修正ルール	編集手順	編集距離
ヵわ ⇒ かは	置換：2 回	4
ヵわ ⇒ か	置換：1 回，削除：1 回	3
ヵわ ⇒ したらいいのか	置換：2 回，挿入：5 回	9
ヵわ ⇒ なっているか	置換：2 回，挿入：4 回	8

がなで表記されるべき語がカタカナで表記されている例が多いことや「ゎ」という文字は「は」と「わ」に置き換えられる可能性が高いなど，挿入，削除，置換されやすい文字に傾向がある．これらの傾向を考慮して，特定の編集のコストを小さく設定することも有効である．本稿における実験では，ブログで頻繁に見られる編集のコスト 100 件をあらかじめ登録することによって，より高精度な修正を実現した．

4.2.3　修正前後の形態素解析コスト値に基づくスコアリング

くだけた表現が出現する文脈における修正ルールの適応度合いを評価する指標として，形態素解析コスト値[14] を用いる．形態素解析コスト値とは単語の生起確率（生起コスト）や単語どうしの連接確率（連接コスト）などから算出される値で，複数ある単語区切り方のうち，尤度の高いものを選択する際に多くの形態素解析器で用いられる．提案手法では，修正箇所周辺における表現の自然さを推定する指標として，形態素解析コスト値を用いる．修正ルールの適用により，不自然な表現が生成された場合，表現周辺の生起コストや連接コストは大きくなることから，修正の誤りを推定する．

図 3 に形態素解析コスト値に基づくスコアの算出例を示す．各形態素における連接コストと単語生起コストの和を文頭からの累積で算出した値（累積コスト）の文末における値が文全体の単語区切りの尤度を表すと考える．くだけた表現を含む「できるかどう ヵわ 分かりません」のような文は未知語部分の単語生起コストが大きいため，文全体の累積コストが大きくなる．修正ルール適用後の文の形態素解析コスト値をそれぞれ算出し，修正前の文の形態素解析コスト値との差分を修正ルールのスコアとする．一般的に短い文の方が形態素解析コスト値は小さくなる傾向にあるが，多数の文字列を削減するような修正ルールは 4.2.2 項の編集距離によるスコア値が低くなる．

4.2.4　総合スコアの算出

修正ルールの総合的なスコア $score$ は修正候補文字列の出現頻度 $freq$，修正文字列間の編集距離 $dist$，形態素解析コスト値の差分 $cost$ を用いて一般的に式 (1) のように記述できる．ここで，関数 f，g，h は各指標の重み付け関数である．本稿における実装では式 (2) のように簡潔にそれぞれ定数 α，β，γ とした．

$$score = f(freq) + g(dist) + h(cost) \quad (1)$$

$$score = \alpha \cdot freq + \beta \cdot dist + \gamma \cdot cost \quad (2)$$

```
修正前の文：　できるかどう ヵわ 分かりません
単語区切り：　できる/ か / どう /ヵわ/分かり/ませ/ ん
累積コスト：　5742/8263/11751/34685/39098/40388/39914
文全体のコスト：39914

修正候補１：　できるかどう かは 分かりません
単語区切り：　できる/ か / どう / か / は /分かり/ませ/ ん
累積コスト：　5742/8263/11751/14430/15438/19341/20631/20157
文全体のコスト：20157　修正前の文との差分（形態素解析コスト）:-19757

修正候補２：　できるかどう か 分かりません
単語区切り：　できる/ か /どうか/分かり/ませ/ ん
累積コスト：　5742/8263/16737/20120/21410/20936
文全体のコスト：20936　修正前の文との差分（形態素解析コスト）:-18978

修正候補３：　できるかどう したらいいのか 分かりません
単語区切り：　できる/か/どう/し/たら/いい/の/か/分かり/ませ/ん
累積コスト：　5742/8263/11751/ ...(略)... /26035/27325/26851
文全体のコスト：26851　修正前の文との差分（形態素解析コスト）:-13063

修正候補４：　できるかどう なっているか 分かりません
単語区切り：　できる/か/どう/なっ/て/いる/か/分かり/ませ/ん
累積コスト：　5742/8263/11751/ ...(略)... /22975/24265/23791
文全体のコスト：23791　修正前の文との差分（形態素解析コスト）:-16123
```

図 3　形態素解析コストを用いたスコアリング例
Fig. 3　Scoring based on morphological analysis cost.

表 4　修正例と各指標におけるスコア（$\alpha = 1$, $\beta = -16$, $\gamma = -0.005$）
Table 4　Integrated scores based on each criterion ($\alpha = 1$, $\beta = -16$, $\gamma = -0.005$).

修正例	出現頻度 (%)	編集距離	形態素解析コスト	総合スコア
てたょ⇒てた	20	1	−15757	83.3
てたょ⇒てたよ	0	2	−14037	38.2
てたょ⇒てたい	2	2	−10946	24.7
てたょ⇒てたといえよう	2	6	−9108	−48.5
今日ゎ午前⇒今日は午前	95	2	−13421	130.1
今日ゎ午前⇒今日午前	0	1	−13131	49.7
今日ゎ午前⇒今日だけ午前	0	3	−6247	−16.76
ぉ金無い⇒金無い	0	1	−13131	49.7
ぉ金無い⇒お金無い	0	2	−10974	22.9
ぉ金無い⇒税金無い	8	2	−9887	25.4
ぉ金無い⇒う金無い	4	2	−6654	5.3

重み付け定数 α，β，γ は，式 (2) における各スコア値が均等に影響し，より多くのくだけた表現が正しく修正されるよう設定する．5 章における実験では，人手によって修正の正解，不正解の判定を行った修正例をもとに，重み付け定数の最適化と，くだけた表現修正における再現率，適合率のトレードオフについて評価する．

表 4 に実際に得られた修正ルールとそのスコアの例を示す（ここでは例として，$\alpha = 1$, $\beta = -16$, $\gamma = -0.005$ とした）．くだけた表現「てたょ」の修正例では出現頻度や形態素解析コストから,「てた」が最適な修正候補文字列であることを示している.「てたといえよう」なども修正候補文字列として得られているが，編集距離が大きいため選択されない.「今日ゎ午前」の修正例では出現頻度により,「今日は午前」が最適な修正であると判定する.「ぉ金無い」の修正例では「税金無い」なども高い総合スコアを得ているが，形態素解析コストが小さい「金無い」が文としてより一般的であると判定する．

4.3　修正ルールの適用と学習

スコアリングした修正ルールの適用と学習について説明する．修正ルールはスコアリングされており，適用する閾値を設定することが可能である．閾値を低く設定した場合，より多くのくだけた表現を修正できる（再現率が向上する）が，修正ルールの誤適用も増加する（適合率が低下する）．閾値を高く設定した場合，再現率は低下するが，適合率は向上する．スコアリングの各指標の重み付け定数 α，β，γ と適用閾値 θ の最適化とくだけた表現修正の再現率，適合率のトレードオフについては 5 章で評価する．

また，提案手法では閾値以上のスコアを持つ修正ルールをデータベースに登録することで，修正ルールの再利用も行う．複雑なくだけた表現など，有効な修正候補が得られない場合に，データベースに登録された修正ルールを参照することで，有効な修正ルールを取得することができる．たとえば,「ゎナニ∪は」（「わたしは」と読む）のような複雑なくだけた表現を修正する際，検索クエリが「*は」のようになってしまい，有効な修正候補が得られない場合がある．このとき，他の文例からすでに「ゎ⇒わ」や「∪⇒し」の修正ルールが得られていれば,「わナニしは」と修正した後,「わ*しは」をクエリとすることで,「わたしは」という正しい表現に修正することができる．同時にこの文例から,「ナニ⇒た」という新たな修正ルールを獲得する．

5.　性能評価実験

提案手法を実装し，性能評価実験を行った．初めに，提案手法の修正ルールをスコアリングする各指標の重み付けパラメータ α，β，γ と修正ルールを適用する閾値 θ を調整することで，くだけた表現修正の再現率と適合率のトレードオフについて評価した．次に，従来手法である，人手による**辞書拡張手法**[4] と修正ルールの**統計的スコアリング手法**[5] の 2 手法の性能と提案手法の性能を比較評価した．辞書拡張手法における課題としては辞書拡張ルールの過剰適用による単語区切りの誤りがあげられる．統計的スコアリング手法における課題としては初期ルールを人手により与える必要があるため，修正可能な未知語数が少ないことがあげられる．これらの課題を考慮して，性能比較評価実験では未知語の出現率や単語区切りの正しさについて評価した．

5.1　提案手法に対する評価実験

5.1.1　実験の手順と環境

提案手法の性能を評価するため，初めに文章に出現する未知語のうち，正規な表現から派生したくだけた表現と，未知語であるがくだけた表現ではないもの（固有名詞や擬音語など形態素解析辞書に未登録のもの）の割合を人手により判定する．具体的には，対象文章を形態素解析器によって形態素解析し，未知語と判定されたものについて，人手によりくだけた表現であるかどうかを判定する．文章の総形態素数，未知語数，くだけた表現数を求める．

次に，再現率，適合率を以下のように定義し，重み付け関数 α，β，γ および，修正ルール適用閾値 θ を，人手によって正しい修正が付与された学習用文章を用いて調整することで，再現率と適合率のトレードオフを評価する．

- 再現率＝正しく修正されたくだけた表現数/対象文章中のまったくだけた表現数
- 適合率＝正しく修正されたくだけた表現数/提案手法で修正したくだけた表現数

表 5　意味変化の大小による修正の分類評価例
Table 5　Categorization of substitution rules based on their effects on the meanings.

修正前	修正後
(a) 意味変化が小さい修正例	
今日は猫ちゃん来てたょー	今日は猫ちゃん来てた
とぉっても気持ちいい	とっても気持ちいい
おいしそぉだったんだけどね	おいしそうだったんだけどね
めッちゃ汗かいた	めっちゃ汗かいた
(b) 意味変化が大きい修正例	
じゃあ，② 時に駅前で	じゃあ，七時に駅前で
可愛いすぎぃ	可愛すぎない
おっはよぉー	おっはよい
(c) 意味変化の有無が判定しにくい修正例	
来ぉへんよっ	来へんよ
私も遅くなる時ぁる∪	私も遅くなる時もある
ぜひぉためしあれ	ぜひためしあれ

トレードオフを確認するため，いくつかの再現率を設定し，各再現率において適合率が最大となるようなパラメータ（α，β，γ，θ）の組を探索する．これは制約付き最適化問題と考えることができ，一般的に $\alpha = 1$ と固定すると，次式のように表すことができる．

$$max.\quad Precision(\beta, \gamma, \theta)$$
$$s.t.\quad Recall(\beta, \gamma, \theta) \geq R \qquad (3)$$

ここで，Precision，Recall はそれぞれパラメータ β，γ，θ によって定まる適合率，再現率である．また，R は設定する再現率である．実験では，3 つのパラメータの値をそれぞれ異なる粒度で変化させることにより，網羅的に解を探索した．

また，修正の正しさについては，各くだけた表現を下記の 3 つの評価基準を用いて人手により分類した．**(a)** 修正前後で意味はほとんど変化していない，**(b)** 修正前後で意味が明らかに変化している，または文の意味が理解できない，**(c)** (a)，(b) の判断がつかない．各評価基準に分類される修正の例を**表 5** に示す．(a) の意味変化が小さいと分類された修正では，与える印象はわずかに変化するかもしれないが，内容や事実関係に変化はないと考えられる．(b) の意味変化が大きいと分類された修正では，文の内容や事実関係に変化があったり，文が意味をなしていなかったりする．(c) の意味変化の有無が判定しにくい修正例では，文として不自然さはあるが，修正前と同じ意味ととらえられるものや，前後の文脈によっては意味が変わりうる修正などが含まれる．意味変化率の算出では (b) と (c) に分類される例を意味が変化したと判定する．

以下に実験環境の詳細を示す．形態素解析器は MeCab [3] を用いた．ある程度チューニングされた形態素解析辞書を想定し，標準 IPADIC 辞書に新聞文書等で頻出する未登録の単語 18 万語を追加登録した拡張 IPADIC 辞書（**基本辞書**）を用いた．

- 形態素解析器：MeCab Version 0.97
- 形態素解析辞書（基本辞書）：MeCab 標準 IPADIC 辞書に新聞文書で頻出の 18 万語を追加
- プログラム実行環境：CPU 2.33 GHz 8 core，RAM 64 GB，OS Linux version 2.6.24, gcc version 4.1.2
- 提案手法の修正候補文字列取得用文書：毎日新聞（2007 年〜2008 年）100 万文
- くだけた表現を含む学習用文章（パラメータ調整用）：商用ブログ 2 万文
- くだけた表現を含む評価対象文章：商用ブログ 5 万文

5.1.2　実 験 結 果

評価対象文章（ブログ 5 万文）中の総形態素数，未知語数，くだけた表現の割合を**図 4** に示す．新聞文書をもとにチューニングされた基本辞書による形態素解析結果では，総形態素数に対する未知語の割合が 1.613% であり，未知語のうちくだけた表現は 60.17% であった．文単位で集計すると未知語を含む文の割合は 12.78% であり，そのうちくだけた表現を含む文は 57.46% であった．

次に，提案手法の性能を評価するため，再現率 10%〜90%に設定したとき，最大の適合率を与えるようなパラメータ（α，β，γ，θ）の組を学習用文章を用いて求めた．同じパラメータにおいて，評価対象文章を修正したときの再現率および適合率の関係と合わせて**図 5** に示す．提案手法では再現率が 70% においても適合率 95% 以上を達成可能であることを確認した．

5.2　提案手法と従来手法の性能比較評価実験

5.2.1　実験の手順と環境

辞書拡張手法，統計的スコアリング手法，提案手法の 3 手法について，評価対象文章（商用ブログ 5 万文）の形態素解析を行い，下記の指標を評価した．(1) 各手法の適用により，単語区切りが向上した未知語の割合（**単語区切り向上率**），(2) 各手法の適用により，単語区切りが悪化した未知語の割合（**単語区切り悪化率**），(3) 各手法の適用により，意味が変化した未知語の割合（**意味変化率**），(4) 評価対象文章全体の形態素数に対する未知語数の割合（**未知語出現率**）．(1)，(2)，(4) は辞書拡張手法，統計的スコアリング手法と提案手法の性能を比較するため，(3) は統計的スコアリング手法と提案手法の性能を比較するために行

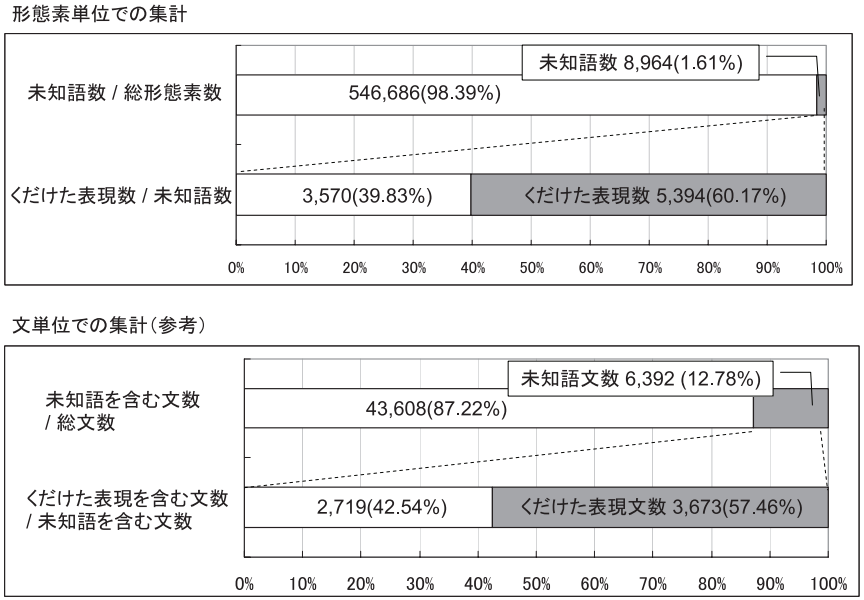

図 4　評価対象文章におけるくだけた表現の割合
Fig. 4　Ratio of peculiar expressions appeared in the blog documents.

う．辞書拡張手法では，文章の表層を変化させないため，本稿で定義した表層上の意味変化は起こらない．

単語区切りの向上とは，基本辞書を用いた形態素解析において単語区切りに誤りを含む文が各手法によって正しく区切られる場合を指す．反対に，基本辞書では正しく単語区切りが行われていた文が各手法によって誤った単語区切りが行われた場合を悪化とする．単語区切りの正解判定は文献 4)，5) と同様に，文献 15) の手法を用いた．具体的には，人手で付与した正解の単語区切りと各手法における形態素解析結果の単語区切りを比較する．提案手法と統計的スコアリング手法では修正により，表層が変化するが，修正後の文の単語区切りが正しければ正解とする．上記の (1)〜(3) については，それぞれの手法で単語区切りまたは表層に変化のあった文のうち 600 文をサンプリングし，評価を行った．

実験環境は 5.1.1 項と同様とした．辞書拡張手法を評価するため，文献 4) を参考に辞書拡張ルールを作成し，機械的に形態素解析辞書に反映させた．修正ルールの統計的スコアリング手法では人手により与えた 250 件の修正ルールをもとに学習用ブログデータ 1,000 万文を用いて学習を行い，修正ルールをスコアリングした．

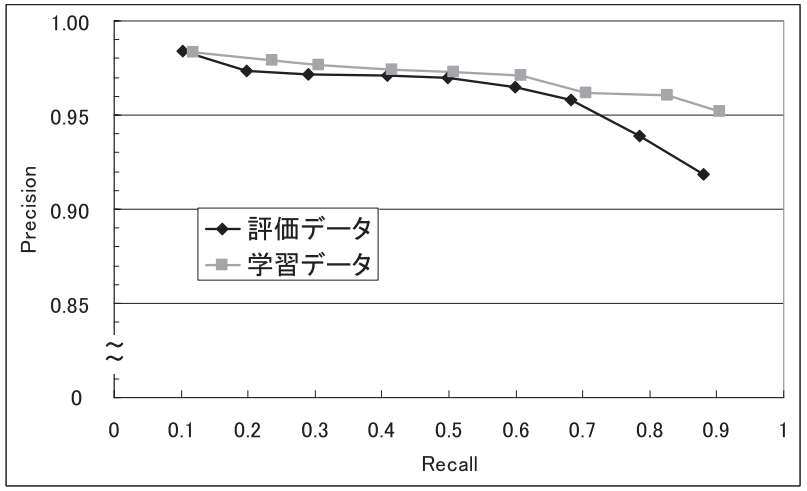

図 5　提案手法における再現率と適合率の関係
Fig. 5　Relation between the recall rate and the precision rate of the proposed algorithm.

5.2.2　実 験 結 果

提案手法，辞書拡張手法，統計的スコアリング手法それぞれにおける単語区切りの向上率と悪化率，意味変化率，未知語出現率を**表 6** に示す．ここで，人手により初期ルールを与える統計的スコアリング手法と同程度の単語区切りの向上率と悪化率，意味変化率となるように提案手法におけるパラメータを調整した．提案手法は辞書拡張手法と比べると単語区切り悪化率，未知語出現率共に低い．ここで，各手法の適用により，基本辞書から未知語が減少した度合いを**未知語減少率**とし，次のように定義する．

- 未知語減少率 ＝ 1 − 各手法における未知語出現率/基本辞書における未知語出現率

提案手法における未知語減少率は 36.1% であり，統計的スコアリング手法の 14.4% と比べると，2.5 倍以上である．提案手法では大規模な修正ルール集合を自動的に構築し，3 つの指標を用いて修正ルールをスコアリングすることで，高精度な修正ルールの適用を可能とする．これにより，従来手法と同程度の修正の正しさを維持しながら，より多くの未知語を削減することが可能となる．

表 6　各手法の性能比較
Table 6　Performance evaluation of each algorithm.

手法	単語区切り向上率 (%)	単語区切り悪化率 (%)	意味変化率 (%)	未知語出現率 (%)
基本辞書	–	–	–	1.613
辞書拡張手法	48.4	29.6	–	1.492
統計的スコアリング手法	51.7	9.3	4.1	1.381
提案手法	52.2	8.4	3.8	1.031

具体的には，基本辞書における未知語出現率（数）は 1.63%（8,964 個）であり，そのうちくだけた表現は 60.17%（5,394 個），それ以外の未知語は 39.83%（3,463 個）であった．提案手法の適用により，くだけた表現 3,344 個とくだけた表現でない未知語 107 個を修正したため，それぞれの比率は 37.18%（2,050 個），62.82%（3,463 個）となった．くだけた表現の修正の再現率，適合率はともに 100% ではないため，未知語中のくだけた表現のみを識別することは困難であるが，提案手法でくだけた表現の大半を解決した後に，残りのくだけた表現以外を多く含む未知語を人手により形態素解析辞書に登録することで，作業の効率化を図ることができる．

これらの実験結果から，提案手法は従来手法の課題であったルールの過剰適用やスケーラビリティの少なさといった問題を解決することができることを確認した．また，計算時間についても従来の統計的スコアリング手法はブログ 1,000 万文による事前学習に 17 時間を要するのに対し，提案手法はパラメータ調整以外の事前学習を必要としない．修正に要する時間は提案手法で新聞コーパス量が 100 万文のとき，入力されたブログ文章 1,000 文を約 1 秒で修正することができ，これは従来手法と同程度の計算時間である．

6. ま と め

本稿ではブログ上の文書に頻繁に見られる口語的な表現や特有の表記などのくだけた表現を正規な表現へ修正する手法を提案した．提案手法では，くだけた表現の修正候補文字列をくだけた表現の少ない文書から自動的に検索することで，正規な表現へと修正するための修正ルールを生成する．生成した修正ルールを修正候補文字列の出現頻度，修正文字列間の編集距離，形態素解析コストの 3 つの指標を用いてスコアリングすることで，最適な修正ルールを適用することができる．

性能評価実験では，修正ルールをスコアリングする 3 つの指標の重み付けパラメータと修正ルールを適用するスコアの閾値を調整することで，くだけた表現修正の再現率と適合率のトレードオフについて評価した．加えて，従来手法である辞書拡張手法，統計的スコアリング手法との性能比較評価実験を行った．提案手法は，くだけた表現の少ない文書から大規模な修正ルール集合を自動的に構築するとともに，3 つの言語的な指標を用いて高精度に修正ルールを適用することで，従来手法と同程度の単語区切りの正確さを維持しながら，従来手法の 2.5 倍以上の未知語を解消できることを確認した．提案手法による未知語の減少率は対象文章に出現する未知語の 36.1% に相当する．

謝辞　本研究は，(独）情報通信研究機構の委託研究「高度通信・放送研究開発委託研究/インターネット上の違法・有害情報の検出技術の研究開発」の一環として実施いたしました．また，日頃ご指導いただく KDDI 研究所伊藤泰彦会長，秋葉重幸所長，松本修一副所長，および菅谷史昭執行役員に深謝いたします．

参 考 文 献

1) 中島伸介，稲垣陽一，草野奉章：高信頼性情報の提示を目指した熟知度に基づくブログランキング方式の提案，日本データベース学会論文誌，Vol.7, No.1, pp.257–262 (2008).
2) 関口裕一郎，川島晴美，奥田英範，奥　雅博：ブログ発信者の特徴を利用した話題抽出手法，日本データベース学会論文誌，Vol.5, No.1, pp.9–12 (2006).
3) Kudo, T.: Mecab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourceforge.net/
4) 風間淳一，光石　豊，牧野貴樹，鳥澤健太郎，松田晃一，辻井潤一：チャットのための日本語形態素解析，言語処理学会第五回年次大会発表論文集，pp.509–512 (1999).
5) 池田和史，柳原　正，松本一則，滝嶋康弘：ブログ的表記を正規化するためのルール自動生成方式の提案と評価，日本データベース学会論文誌，Vol.8, No.1, pp.23–28 (2009).
6) 竹元義美，福島俊一：口語的表現を含む日本語文の形態素解析の実現と評価，情報処理学会自然言語処理研究会報告，pp.105–112 (1994).
7) 竹下　敦，福永博信：話し言葉に対する形態素解析，情報処理学会第 42 回全国大会，pp.1C–3 (1991).
8) 松本裕治，伝　康晴：話し言葉の形態素解析，情報処理学会音声言語情報処理研究会報告，pp.9–14 (2001).
9) Masuyama, T., Sekine, S. and Nakagawa, H.: Automatic Construction of Japanese KATAKANA Variant List from Large Corpus, *Proc. 20th International Conference on Computational Linguistics* (*COLING*), pp.1214–1219 (2004).
10) Murawaki, Y. and Kurohashi, S.: Online Acquisition of Japanese Unknown Morphemes using Morphological Constraints, *Proc. 2008 Conference on Empirical Methods in Natural Language Processing* (*EMNLP-2008*), pp.429–437 (2008).
11) Mori, S. and Nagao, M.: Word extraction from corpora and its part-of-speech

estimation using distributional analysis, *Proc. 11th International Conference on Computational Linguistics* (*COLING*), pp.1119–1122 (1996).
12) 柳原　正，松本一則，池田和史，滝嶋康弘：情報量によるモデル検定を用いた単語境界推定方式の提案，情報処理学会第 190 回自然言語処理研究会論文集，pp.43–47 (2009).
13) Levenshtein, V.I.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals, *Journal of Soviet Physics, Doklady*, pp.707–710 (1966).
14) Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proc. 2004 Conference on Empirical Methods in Natural Language Processing* (*EMNLP-2004*), pp.230–237 (2004).
15) Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A N-Best Search Algorithm, *Proc. 15th International Conference on Computational Linguistics* (*COLING*), pp.201–207 (1994).

(平成 22 年 3 月 19 日受付)
(平成 22 年 7 月 5 日採録)

(担当編集委員　相良 毅)

池田　和史（正会員）

2006 年大阪大学基礎工学部情報科飛び級のため中退．2008 年同大学大学院博士前期課程修了．同年 KDDI（株）入社，研究所所属．自然言語処理等の研究に従事．電子情報通信学会，日本データベース学会各会員．

柳原　　正（正会員）

2002 年慶應義塾大学環境情報学部卒業．2004 年同大学大学院修士課程修了．2005 年 KDDI（株）入社，研究所所属．リコメンダシステム，テキストマイニング等の研究に従事．電子情報通信学会，日本データベース学会各会員．

松本　一則

1984 年京都大学工学部情報工学科卒業．1986 年同大学大学院修士課程修了．同年国際電信電話（株）入社，研究所所属．現在，KDDI 研究所知能メディアグループにて，マルチメディア検索，コンテンツ配信の研究開発に従事．電子情報通信学会，日本データベース学会各会員．工学博士．

滝嶋　康弘

1986 年東京大学工学部電気工学科卒業．1988 年同大学大学院電子工学専攻修士課程修了．同年国際電信電話（株）（現 KDDI（株））入社．現在，（株）KDDI 研究所知能メディアグループリーダ．この間，動画像の符号化方式，動画通信システム，情報理論の研究・開発に従事．電子情報通信学会，映像情報メディア学会，画像電子学会各会員．工学博士．