崩れ表記語の生成確率を用いた表記正規化と形態素解析

斉藤 いつみ　貞光 九月　浅野 久子　松尾 義博
NTT メディアインテリジェンス研究所
{saito.itsumi, sadamitsu.kugatsu, asano.hisako,
matsuo.yoshihiro}@lab.ntt.co.jp

1　はじめに

近年 Twitter 等を代表とするマイクロブログが普及し，個人によって書かれたテキストを対象とした評判分析や要望抽出，興味推定に基づく情報提供など個人単位のマーケティングのニーズが高まっている．一方このようなマイクロブログ上のテキストでは口語調や小文字化，長音化，ひらがな化，カタカナ化など新聞等で用いられる標準的な表記から逸脱した崩れた表記 (以下崩れ表記語と呼ぶ) が多く出現し，新聞等の標準的な日本語に比べ形態素解析誤りが増加する．

これらの崩れ表記語に対し，崩れた表記を辞書に存在する語（以下正規語と呼ぶ）にマッピングして解析を行うという表記正規化の概念に基づく解析が複数提案され，有効性が確認されている [1, 2, 5]．日本語における表記正規化と形態素解析手法としては，大きく (1) ルールにもとづいて入力文字列の正規化候補を列挙しながら辞書引きを行う方法 [7, 8, 9]，(2) 多数の崩れ表記語を列挙し，列挙した形態素のコストをモデルによって学習する方法 [3, 6, 10] が存在する．(1) では，事前に人手で定めた文字列レベルのルールに基づき，崩れた文字列に対し正規文字列を展開しながら解析するシンプルな方法が提案されている (例えば,「ぉ→う」という文字列正規化ルールを作成し，“ぉ”という文字列が入力文に含まれる場合 “ぉ”を “う”に変換した文字列に対しても辞書引きを行い形態素ラティスを拡張する)．(2) では，鍛治ら [3] は形態素正解データ，斉藤ら [6] は形態素正解と対応する正規語の正解データを用いて識別モデルを学習し，崩れ表記語を精度よく解析する方法を提案した．

正規語を考慮した形態素解析モデルのコスト学習において，文献 [3, 6] では正解データを用いた識別学習を行っており，学習データに精度よくパラメタチューニングを行うことが可能になる．しかし通常崩れ表記語が出現する新しいドメインの形態素正解データ量（例えば Twitter）は新聞などの広く共有されている言語資源に比べ少量であるため，すべての崩れ表記語に対して適切なコストを学習することは難しい．そのため未知データにも頑健なモデルを学習するためには外部知識（大規模コーパスから計算した単語の出現分布など）を導入することが有効であることが多い．文献 [3] でも，大量の平文を自動解析して構築した言語モデルを学習の素性として導入することが有効であることが示されている．文献 [6] では言語モデルの他に崩れ表記と正規表記のアノテーションデータから求めた文字列変換確率を考慮しているが，この手法は一定量の崩れ表記-正規表記ペアの人手アノテーションを前提としており，データの作成コストが発生する．また，文字列変換確率では単語ごとに異なる崩れ語の出現確率を考慮することができない．

表 1: 崩れ表記語の分類と本研究の対象範囲

大分類	小分類	具体例
口語	促音の挿入，置換	かっわいい, いこっか
	長音の挿入，置換	ねむーーい, とーきょー
	母音の挿入	まあるく, きたあああああ
	発音の崩れ	すっげぇ, くだしゃい
異表記	小文字化	いぃよ,おぅち
	カタカナ/ひらがな化	アリガトウ, てすと
	同音異表記	まぢ, 少しづつ
	形の類似	ネ申
誤字脱字	タイプミス	これをを
	送り仮名誤り	面倒臭せ
	漢字の誤用	待ち通しい
略語	略語	おめ, よろ
方言	方言	やらへん, してんねや

本研究では，これらの正解付きデータに基づく学習に新たな外部知識を導入することを提案する．具体的には，大量の平文から学習した単語レベルの崩れ確率（正規語から崩れ表記語が生成される確率）を学習し，少量の正解付きデータを用いた識別モデル学習に素性として導入する．この際，崩れ表記語と正規語の列挙については文献 [6] の手法を用いて文字レベル，単語レベルの正規化候補展開を行い，展開された崩れ表記語の生成確率を求める際には [10] の手法を用いる．平文データから学習された単語レベルの崩れ特性を表す素性の導入によって，少量の学習データに出現しない多様な崩れ表記語に対しても崩れ表記語ごとの出現分布を反映することが可能となり，より頑健なモデル学習が可能になることを示す．

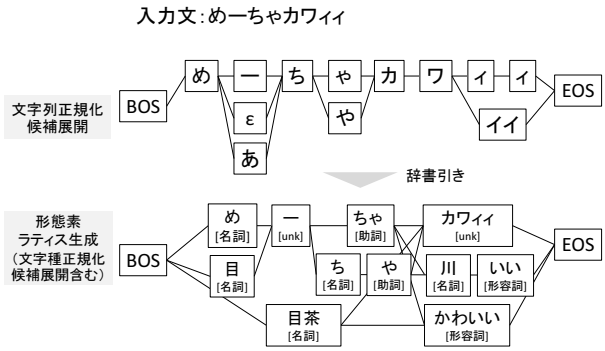

図 1: 文字列正規化候補展開と形態素ラティス生成の例

表 1 には，崩れ表記語の分類と本研究で扱う範囲（網掛け部）を示した．対象範囲は，音的な類似という点で特定のパタンが存在すると考えられる口語調の崩れ表記や，異表記（小文字化，同音異表記，ひらがな化，カタカナ化）とした．これらを対象とした理由は，崩れ表記語全体の中で占める割合が大きく今回の提案手法で統一的に表現できる現象であったためである．

本研究の構成は次の通りである．2 章で提案手法（正規語の展開，崩れ表記語生成確率を含む生成モデルの推定，形態素解析・表記正規化モデル学習）について詳述し．3 章で実験内容について示す，4 章でまとめと今後の課題を示す．

2 提案手法

本章では，提案手法の全体について述べる．提案手法は主に，正規語候補の列挙と形態素ラティスの構築方法，新たな素性として用いる生成モデルの学習，表記正規化と形態素解析のためのモデル学習の 3 つからなり，それぞれ 2.1 節, 2.2 節, 2.3 節で詳述する．

2.1 正規表記の列挙と形態素ラティス構築

形態素ラティスの構築は，基本的には辞書を用いて行う．ただしこの際，辞書に存在しない崩れ表記語を列挙するため，文献 [6] で提案された手法を元に崩れ表記語と正規語の列挙を行った．具体的には，予め学習データから求めた文字レベルの表記正規化パタンに基づく文字レベルの正規化と文字種レベル（ひらがな表記，カタカナ表記）の正規化を考慮して崩れ表記語と正規語の候補ペアを列挙する方法である．この際，[6] と同様に事前正規化として,「っ」と「ー」の連続に関しては 1 文字まで縮約させ，母音の連続に関しては 3 文字まで縮約させる，という処理を行った．また，文字列正規化については，正解データから獲得されたパタンの他に，既存研究 [9] で提案されているルールなど容易に人手で生成できるルールも適用した．

図 1 に文字列正規化候補展開，文字種正規化候補展開と形態素ラティス生成の例を示す．図 1 に示すように，文字列レベルの可能な変換候補と文字種レベルの可能な変換候補を考慮することにより，“カワィィ”→“かわいい”などの崩れ語・正規語ペアの列挙が可能になる．表 2 に用いた文字列正規化パタンの例を示す．ここで，c_w は崩れ文字列，c_v は正規文字列を表す．

表 2: 獲得した文字列正規化パタン例

c_w	c_v	c_w	c_v
ねー	ない	ヴァ	バ
ねえ	ない	きょ	きょう
っ	い	しゅ	す
にゃ	な	っげー	ごい
ぅ	う	みぃ	むい
おお	う	ー	あ

解析時には正規語と表出表記双方の情報を考慮するため，ラティス構築の際は辞書引きされた正規語と表出表記，正規語の品詞の 3 つ組の情報を各形態素ラティスのノードに保持する．ここで，表出表記とは，正規語の観測された表層形のことを表す．例えば，“カワィィ”という入力表記に対して “かわいい（形容詞）”という辞書エントリが列挙された場合には，“カワィィ，かわいい，形容詞”の情報をノードに保持する．この際の表出表記は “カワィィ”であり，正規語は “かわいい”である．また，すべての可能な候補を考慮すると形態素ラティスが大きくなり計算量が膨大となるため，入力文の文字列を正規表記に置き換えた場合と置き換えない場合の文字 n-gram 素性を用いて，閾値以下の候補については枝刈りを行った．

2.2 形態素正解なしデータを用いた崩れ表記語生成確率の学習

本節では，2.3 節で述べる形態素解析・表記正規化のモデル推定の素性として用いる崩れ表記語生成確率の推定について述べる．冒頭で述べたように，今回素性として用いる崩れ表記語生成確率の推定は形態素正解がアノテーションされていない平文コーパスを用い，正解付きデータに比べ大規模なコーパスにおける表出表記の出現分布を学習する．この際，2.1 節で示した正規語の列挙方法を用いて大量の崩れ表記語-正規語ペアを生成し，多様な崩れ表記語に対して生成確率を推定した．学習には，[10] で提案された生成モデルを用いた．[10] は下記のように単語の生成確率を定義し形態素解析の問題として定式化した．

$$(\hat{\boldsymbol{w}}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{t}}) = \arg\max_{\boldsymbol{w},\boldsymbol{v},\boldsymbol{t} \in S} \prod_{i=1}^{n} P(w_i|v_i)P(v_i|t_i)P(t_i|t_{i-1}).$$

ここで，$\boldsymbol{v} = (v_1, v_2, \ldots v_n)$ と $\boldsymbol{t} = (t_1, t_2, \ldots t_n)$ はそれぞれ潜在的な正規語の系列と品詞の系列である．$\boldsymbol{w} = (w_1, w_2, \ldots w_n)$ は $\boldsymbol{v}$ の観測された表出表記系列，S は入力文に対して可能な $(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{t})$ の集合を表す．$P(w_i|v_i)$, $P(v_i|t_i)$, $P(t_i|t_{i-1})$ は観測できないため，EM アルゴリズムを用いて推定する方法が示され

表 3: 正規語 “すごい（形容詞）”の表出表記とその確率の推定結果（上位 10 位）

表出表記	$p(w\|v)$	表出表記	$p(w\|v)$
すごい	0.623	すんごい	0.028
すげー	0.115	すご	0.013
すっごい	0.069	すっげー	0.010
すげぇ	0.046	スゴい	0.010
すげえ	0.032	すっげ	0.007

ている．[10] では$\boldsymbol{w}$の候補集合で崩れ表記語に関してはひらがな表記を対象としていたが，ひらがな以外の崩れた表記についても適用が可能であるため，本研究でも上記の手法にしたがって推定を行った．

推定結果の例を表 3 に示す．表 3 では上位 10 件のみを示しているが，”すごい（形容詞）”に関しては 8 万文の学習データから約 30 パタンの崩れ表記語生成確率を獲得することができた．

2.3 形態素正解付きデータを用いた形態素解析・表記正規化モデルの学習

2.3.1 形態素解析・表記正規化モデル

本項では，形態素解析と表記正規化のためのモデルについて述べる．本研究では，入力文 s に対し正しい表出表記系列 $\boldsymbol{w} = (w_1, w_2, \dots w_n)$，正規語系列 $\boldsymbol{v} = (v_1, v_2, \dots v_n)$，品詞系列 $\boldsymbol{t} = (t_1, t_2, \dots t_n)$ を求める問題を考える．この問題は次のように定式化できる．$(\hat{\boldsymbol{w}}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{t}}) = \underset{(\boldsymbol{w},\boldsymbol{v},\boldsymbol{t})\in L(s)}{\arg\max}\ \mathbf{w}\cdot\mathbf{f}(\boldsymbol{w},\boldsymbol{v},\boldsymbol{t})$．ここで $(\hat{\boldsymbol{w}}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{t}})$ は最適な系列，$L(s)$ は入力文 s に対し構築される形態素ラティス（各ノードは表出表記，正規表記，品詞の 3 つ組情報を持つ），$\mathbf{w}\cdot\mathbf{f}(\boldsymbol{w},\boldsymbol{v},\boldsymbol{t})$ は重みベクトル $\mathbf{w}$ と素性ベクトル $\mathbf{f}(\boldsymbol{w},\boldsymbol{v},\boldsymbol{t})$ の内積を表す．最適系列は $\mathbf{w}\cdot\mathbf{f}(\boldsymbol{w},\boldsymbol{v},\boldsymbol{t})$ の値にしたがって選択される．形態素解析のためのモデル推定の方法については，正解付きデータを用いた識別学習を行う．本稿では [3] で提案された潜在パーセプトロンを用いて学習を行う．ここで [3] は，表出表記，表出品詞，正規語，正規品詞の 4 つ組で定式化を行っていたが，本研究では表出表記と正規語の品詞はほぼ同一とみなせると考え，表層表記，正規語，品詞の 3 つ組を考慮した．

2.3.2 素性

素性は，[3] や [6] でも用いられている基本素性として，表層表記・正規語・品詞三つ組素性 $f_{w_i v_i t_i}$，品詞 bi-gram 素性 f_{t_{i-1},t_i}，正規語・品詞素性 $f_{v_i t_i}$，大規模データから求めた言語モデル（正規語・品詞 bigram，品詞 bigram）素性 $-\log p_{v_{i-1}t_{i-1},v_i t_i}$，$-\log p_{t_{i-1},t_i}$ を用いる．また本研究の特徴として 2.2 節で求めた生成モデルの値 $p(w|v,t) = p(w|v)\cdot p(v|t)$ を用いる．ここで，表層表記・正規語・品詞三つ組素性，品詞 bi-gram 素性，正規語・品詞素性については，出現する場合に 1，それ以外に 0 となる 2 値変数であり，言語モデル素性と生成モデル素性は連続変数である．ここでの我々の狙いは，生成モデルの推定値を素性として用いることで，正解データのみからは得ることのできないより広いコーパスにおける崩れ表記語の分布をモデルに反映するという点である．

3 実験

3.1 実験データ

本研究では，次の各ステップに際しそれぞれデータを用意した．(1) 2.1 節で示した形態素ラティス生成のための文字列正規化パタン学習，(2) 2.2 節で示した生成モデルの学習，(3) 2.3 節で示したモデル学習に素性として用いる言語モデル（正規語, 品詞の bi-gram モデル），(4) 2.3 節で示し形態素解析・表記正規化モデルの学習，(5) 評価，である．(1) の文字列正規化パタン学習には，人手アノテーションしたブログ，Twitter データから収集した崩れ表記と対応する正規表記のペアデータ約 17000 ペアを用いた．(2) の言語モデルは，ブログの形態素正解ラベルなしデータを約 824 万文を Mecab を用いて自動解析した結果を用いて構築した．(3) の生成モデル学習には Twitter の平文 8 万文を用いた．(4),(5) は，ランダム抽出した Twitter データに対し，URL やハッシュタグ，定期ツイート，意味の読み取れない文を人手で削除し，正しい形態素区切りと品詞，正規語を人手アノテーションしたデータを作成した．学習には 1983 文，テストには 2293 文を用いた．また，辞書は Mecab[4] で学習・配布されている IPA 体系の辞書を用いた．

3.2 評価方法と比較手法

本研究では，評価文に対し次の 4 つの手法に基づく解析結果を比較した．1) Mecab で学習されたコスト・辞書を用いて解析した結果，2) 2.2 節で求めた生成モデルを用いて解析した結果，3) 潜在パーセプトロンのベースモデル（基本素性のみを考慮）で解析した結果，4) 提案手法（潜在パーセプトロンのベースモデルに生成モデル素性を導入）．評価指標は，形態素区切り，品詞の Precision, Recall, F-値を用いた．

3.3 実験結果と考察

表 4 には，単語区切りと品詞の評価結果を示した．まず，Mecab の結果と生成モデルを比較すると，2.2 節で推定した生成モデルのみを用いても Mecab に比べて精度を向上できていることがわかる．このことから，形態素正解のついていない文に対し，崩れ表記語を多数生成して [10] の手法を適用した場合でも崩れ表記語の生成確率を適切に求めることができていると考えられる．

次に，パーセプトロンを用いた識別モデルの結果をみると，ベースモデル，提案モデルともに生成モデル

表 4: テストデータにおける形態素区切り，品詞評価結果

	word seg and POS tag		
method	precision	recall	F 値
(1) Mecab	0.765	0.870	0.814
(2) 生成モデル	0.788	0.855	0.820
(3) パーセプトロン（ベース）	0.820	0.875	0.847
(4) パーセプトロン（提案法）	0.825	0.887	0.855

よりも高い精度となっている．特に，生成モデルの推定結果を用いた提案モデルが最も良い結果となっており，正解なし文から推定した崩れ表記語生成確率を用いることでより頑健なモデルが学習できたことが確認できた．

表 5 には，提案手法による解析結果の改善例を示した．ここで，“/”は形態素の境界を表し，括弧内に推定された正規語と品詞を記述した．入力例 1,2 ベースモデル，提案モデルともに正しく解析できている例である．入力例 3 は提案法が正しく解析できている例である．入力例 3 では，ベースモデルが過剰に正規化を行ってしまっているのに対し，提案モデルでは過剰な正規化を抑えながら正しく解析できていることが分かる．今回の提案のように，多数の崩れ表記語を形態素ラティスに列挙する場合，不要な候補が選択されることによるデグレードを抑えながら正しい候補を選択することが重要になるが，提案素性を用いることで単語単位の崩れ確率を考慮可能になり，より頑健性の高いモデルが構築できたと考えられる．ただし，ベースモデルで正解しているが提案法で探索に失敗している例も散見され，さらなるモデルの改良を検討していく必要がある．

表 5: 提案手法による解析結果例

	解析結果例：表出表記 (正規表記, 品詞)
入力例 1	ちょと寒いーなー
Mecab	ちょ(名詞)/と (助詞)/寒い (形容詞)/ー (助詞)/ なー (助詞)/
ベース	ちょと (ちょっと, 副詞)/寒いー (寒い, 形容詞)/ なー (なー, 助詞)/
提案法	ちょと (ちょっと, 副詞)/寒いー (寒い, 形容詞)/ なー (なー, 助詞)/
入力例 2	楽しもっ
Mecab	楽し (形容詞)/もっ(形容詞)/
ベース	楽しも (楽しも, 動詞)/っ(う, 助動詞)/
提案法	楽しも (楽しも, 動詞)/っ(う, 助動詞)/
入力例 3	ゆうたことに責任もとう
Mecab	ゆう (動詞)/た (助動詞)/こと (助動詞)/に (助詞)/ 責任 (名詞)/も (助詞)/とう (動詞)
ベース	ゆう (ゆ, 名詞)/た (た, 助動詞)/こと (こと, 名詞)/ に (に, 助詞)/責任 (責任, 名詞)/もとう (もと, 名詞)
提案法	ゆう (いう, 動詞)/た (た, 助動詞)/こと (こと, 名詞)/ に (に, 助詞)/責任 (責任, 名詞)/もと (動詞)/う (助動詞)

4 おわりに

本研究では，崩れ表記の正規化と解析において形態素の正解付きデータと正解なしデータの双方を用いて，単独で用いるモデルよりも頑健なモデルを学習できることを示した．今後の課題としては，より効果的な素性の検討，正解なしデータを形態素解析のモデル学習に素性としてではなく学習データとして直接用いる方法の検討，正規語候補の効率的な列挙方法の検討などを行っていく予定である．

参考文献

[1] Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a #twitter. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pp. 368–378, 2011.

[2] Bo Han, Paul Cook, and Timothy Baldwin. Automatically constructing a normalisation dictionary for microblogs. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 421–432, 2012.

[3] Nobuhiro Kaji and Masaru Kitsuregawa. Accurate word segmentation and pos tagging for japanese microblogs: Corpus annotation and joint modeling with lexical normalization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 99–109, Doha, Qatar, October 2014. Association for Computational Linguistics.

[4] T. KUDO. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.net/*, 2005.

[5] Chen Li and Yang Liu. Improving text normalization using character-blocks based models and system combination. *Proceedings of COLING 2012*, pp. 1587–1602, 2012.

[6] Itsumi Saito, Kugatsu Sadamitsu, Hisako Asano, and Yoshihiro Matsuo. Morphological analysis for japanese noisy text based on character-level and word-level normalization. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1773–1782, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[7] Ryohei Sasano, Sadao Kurohashi, and Manabu Okumura. A simple approach to unknown word processing in japanese morphological analysis. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 162–170, 2013.

[8] 岡照晃, 小町守, 小木曽智信, 松本裕治. 表記のバリエーションを考慮した近代日本語の形態素解析. 人工知能学会全国大会講演集, 2013.

[9] 勝木健太, 笹野遼平, 河原大輔, 黒橋禎夫. 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. 自然言語処理学会年次大会講演集, pp. 1003–1006, 2011.

[10] 工藤拓, 市川宙, David Talbot, 賀沢秀人. web 上のひらがな交じり文に頑健な形態素解析. 自然言語処理学会年次大会講演集, 2012.