交通行動に関するブログテキストの表記変化の分析と形態素辞書のカスタマイズ

鷹尾 和享
システム科学研究所

1　はじめに

近年では Web 上に多くのブログが発信されており、一般の人々の率直な「生の声」あるいはセンチメントが含まれていることが期待できる。そこで、筆者らは、ブログに書かれやすい話題として、関西空港の利用行動に着目し、交通行動における生の声の分析を試みており[1]、そのためのデータ収集を行った[2]。

ところが、従来の形態素解析システムは、十分に校正のなされたフォーマルなテキストに基づいて開発されている場合が多く、ブログのような砕けたテキストに対しては必ずしも良好な性能を発揮しない。さらに、多くの形態素解析システムがコーパスを使って構築されていることも一因となり、異なる分野に関するテキストへの適用性は十分に考慮されているとは言いがたい[3][4]。

ブログテキストの特徴の一つに表記の変化が挙げられる。たとえば、形容詞をカタカナで記述したり、形態素の途中に長音を挿入したりする場合が見られ、これらの箇所で解析誤りを引き起こすことが多い。これに対処するために、通常の方法では、対象分野のコーパスを用意して再学習させることになるが､それには膨大なコストがかかる。また、形態素辞書へ単語を追加する方法が考えられるが、それにはコスト値という、馴染みのない値を設定する必要がある。そこで、本稿では、既存の形態素辞書の語とそのコスト値を流用し、表記変化したエントリを追加する方法を考えた。

筆者の目的は交通行動に関する生の声の分析であり、形態素解析についてはエンドユーザの立場である。したがって、本稿では、手軽に対処することを第一に考え、ブログテキストの表記変化のパターンを分析し、既存の形態素辞書をカスタマイズする方法を試みたので、報告する。

2　関連研究

形態素の表記変化に対処する試みはいくつか行われている。橋本ら(2011)[5]は長音記号などの口語的表現を扱ったコーパス構築について報告している。また、勝木ら(2011)[6]はオノマトペの小文字化・長音化に対処した形態素解析について報告している。また、齋藤ら(2011)[7]は誤字脱字や伏せ字に対処する方法について報告している。

既存の形態素辞書にも部分的に対応がなされている。MeCab（標準の IPA 辞書）[8]はコーパスから構築されているが、コーパスに登場したものについて収録がなされているようである。また、UniDic[9]はエ音便や促音便については積極的に収録した辞書を提供している。

このように、いくつかの研究はなされているものの、エンドユーザの立場で包括的に「使える」ようにする研究例は必ずしも十分とは言えない。

3　表記変化の分析と形態素辞書のカスタマイズ

3.1 表記変化のパターン分析とカスタマイズの方針

本稿では代表的な形態素解析システムの一つである MeCab[8]を用いてその形態素辞書のカスタマイズを試みる。筆者の先行研究[2]で収集したブログをプレーンテキストに変換し、MeCab で未知語処理なしで形態素解析をさせると、ひらがな1文字で不自然に区切れていたり、明らかに使用頻度の少ないと思われる語として解析されたりしている箇所がかなりあることがわかった｡特に、ひらがなの小文字や長音記号が目立つ。このような場合は、従来ではフィラーとして扱い、フォーマルな日本語の部分だけを抽出する方法で対処さ

れることがあった。しかし、「無理かなーと思って」のように文の途中に登場した場合には語の一部として認識しないと形態素解析に影響すると思われるためと、「長ーい」のように形態素の途中に出現する場合は必然的に長音を含んだ形態素として対処する必要があるため、本稿では形態素の表記変化として扱う方針を採った。

以下、表記変化を整理し、MeCab 辞書をカスタマイズするという視点で説明する。ブログテキストに登場した語だけでなく、網羅的に表記変化を補うことを狙って、変化のパターン別にまとめた。例で示す語の矢印の左側がオリジナルの MeCab 辞書に収録されている形、右側が変化形である。

3.2 カタカナ化

(a)形態素全体

もともとがひらがな表記である形態素全体がカタカナ化する場合は形容詞、動詞の基本形や、代名詞、副詞などに見られる。また、漢字のひらがな化、カタカナ化も見られる。これは、ひらがなの並びに強調したい語が埋没するのを防いだり、漢字表記にした場合の堅苦しさを回避する狙いがあると思われる。

例：かわいい → カワイイ
　　むかつく → ムカツク
　　あなた → アナタ
　　ぐっすり → グッスリ

(b)語幹のみ

語幹のみがカタカナ化する場合があり、形容詞・アウオ段、動詞の五段・ラ行、一段に見られる。この変化は全活用形に亘る。

例：やばい → ヤバい
　　けちる → ケチる
　　ばてる → バテる

動詞に関しては一部の語に限られるとみられるが、その区別が困難なため、本稿では一括して変化形を登録した。また、連用形が名詞化する場合もこのパターンで対処できる。

例：かんじ → カンジ

(c)誤記対策・その逆

これは副詞「○○っと」の「っ」をカタカナで表記するかどうかの違いである。

例：ドバっと → ドバッと
　　ホロッと → ホロっと

3.3 長音等の挿入

(a)形容詞の語尾の先頭

形容詞の語尾の先頭への長音挿入は、話し言葉で語を強調する意味で間を延ばす口調を文字で表現する狙いがあると思われる。また、長音記号ではなく「〜」で表記されることもしばしばある。この変化は基本形、連用テ接続、連用タ接続にみられる。また、助動詞の特殊・タイと特殊・ナイも同様の変化をする。

例：赤い → 赤ーい, 赤〜い
　　赤く → 赤ーく, 赤〜く
　　赤かっ → 赤ーかっ, 赤〜かっ
　　たい → たーい, た〜い

(b)形容詞の２音目

形容詞の２音目に長音や促音が挿入される場合があり、同様に、話し言葉における強調の口調を文字で表現する狙いがあるとみられる。

例：すごい → すっごい
　　すばらしい → すーばらしい

(c)形容詞の末尾

形容詞の末尾に挿入する場合は、余韻を持たせて語への注目を狙う口調を表すものと思われる。

例：赤い → 赤いー

(d)助動詞・助詞の末尾

助動詞「た」「だ」「や」「じゃ」や終助詞等を延ばし、余韻を持たせる狙いがあると見られる。

例：た → たぁ, たー, た〜
　　ぞ → ぞー, ぞぉ, ぞっ

(e)語の途中

副詞などでは語の途中、促音や撥音の直前に長音等が挿入される場合がある。

例：とっても → とーっても，とぉーっても
　　あんな → あーんな

また、最後から２音目や、反復型オノマトペの２音目に長音や促音が挿入されることがある。

例：たまに → たまーに, たまぁに
　　むかし → むかーし
　　ごくごく → ごっくごく

3.4 引き音母音の長音化、小文字化

引き音母音となる場合に、長音や小文字に変化する場合がある。これは表記をあえて砕けさせる狙いがあるとみられる。これは、動詞の五段・ワ行促音便のほか、名詞-形容動詞語幹、副詞、感動詞などにも見られる。

例：ゆう → ゆー, ゆ～
　　きれい → きれー
　　しいんと → しぃんと

3.5 形容詞語尾の脱落＋引き音/促音

形容詞の語尾が脱落し、長音や小文字で引っ張ったり、促音を挿入したりする変化が見られる。

例：おいしい → おいしー, おいしっ
　　甘い → 甘ぁ

3.6 形容詞のエ音便化

形容詞の語幹末尾がア段,オ段,イの場合に基本形がエ音便化する変化が見られる。

例：すごい → すげえ, すげぇ, すげー

3.7 副詞（擬態語「りと」）の促音化

副詞「○○りと」が促音化する場合がある。これは、擬態語に限った変化のようである。

例：ずらりと → ずらっと

3.8 特殊・デス、特殊・マス

助動詞「です」「ます」の途中への長音・促音の挿入および未然形のハ行化が見られる。

例：です → でーす, で～す, でっす
　　ませ → まへ

3.9 衍字または活用規則の誤解

衍字なのか、あるいは活用規則を誤解しているのか定かでないが、形容詞の連用形に「い」が余分に入る場合がある。

例：おいしかっ → おいしいかっ

3.10 組み合わせ

以上の変化パターンが複数組み合わさった場合もある。次の例はカタカナ化と長音挿入の組み合わせである。

例：がらんと → ガランと → ガラ～ンと

3.11 ンカッ型

関西地方の方言に助動詞「ん」「へん」「ひん」があるが、これらがきちんと収録されていないので、追加した。オリジナルの辞書では「あかん」１語のものおよび不変化型「ん」が収録されているが、「あか|し|まへ|ん」のようにもできることから、「あか|ん」と区切るのが正しい。活用形は表１の通りである。

表１：ンカッ型の活用

活用形	活用語尾
基本形	ん, ーん, ～ん
連用タ接続	んかっ
連用テ接続	んかっ, んく, んくっ

3.12 対処が困難なもの

本稿では、コスト値については、変化の元となったオリジナルの形態素のコスト値をそのまま流用することを基本とした。しかし、場合によってはコスト値を調整しないと正しく解析できないものもあった。

また、長音等の挿入はいくつでも可能なので、きりがなく、形態素辞書に追加する方法では実際的には限界があるといえる。

また、本稿で扱わなかった表記変化のパターンには次のようなものがある。

・部分的に漢字→ひらがな
　例：あっと言う間に → あっと言うまに
・ひらがな/カタカナ→漢字
　例：ホタテ → 帆立
・複数の形態素で連語化、カタカナ化
　例：いけ|ない → イケナイ
・その他の挿入、変化
　例：秘密 → ヒ・ミ・ツ

また、表記の変化ではない、通常の語のカバレッジについては本稿では扱っていない。

４　実験

本稿の方法の性能を見るため、実験を行った。ブログテキストからカタカナ・長音（および'〜'）・小文字ぁぃぅぇぉの並びを取り出し、そのうちオリジナルの辞書に名詞-一般、名詞-サ変接続、名詞-固有名詞として収録済みのものを除外し、その中からランダムに 100 個を取り出し、元のテキストの該当箇所をマーキングした。こうすることで、本稿で扱ったような表記変化の箇所や、カタカナ表記の語を中心に抽出される。そして、表記変化を補充した辞書を用いて形態素解析を行い、その箇所が正しく解析されているかどうかをチェックした。結果を表２に示す。

「通常語－もともと収録」「通常語－未知語処理で正しく解析」は表記変化でないカタカナ並びの語や複合語等で正しく解析されたものである。また、「通常語－その他の未知語」とは表記変化ではない通常の語に関するカバレッジ不足の問題であり、感動詞（例：あちゃ〜っ）・固有名詞（例：アンカレッジ）・擬音語（例：ドンパチ）・その他の名詞（例：ポストカード）があった。

これら本稿で問題とするものではないものを除くと、表記変化に関してはかなり対処できていることがわかる（「表記変化－本稿で扱ったもの」）。

「表記変化－本稿で扱わなかった」とは、(i)長音や小文字の挿入個数にきりがない場合（例：がんばぁぁれ）、(ii)一部のみひらがな→カタカナ（例：よめはン）、(iii)外来語の表記の違い（例：エキスプレス）に分類できる。

表２：実験結果

正否	分類	度数
○	表記変化－本稿で扱ったもの	22*
○	通常語－もともと収録	21
○	通常語－未知語処理で正しく解析	37
×	表記変化－本稿で扱わなかった	4
×	ハンドル名、子供の呼称等	4
×	通常語－その他の未知語	11
×	通常語－既知語だが解析誤り	1

* うち、もともと収録されている語でもあるもの: 7

５　結論

本稿では、ブログテキストの砕けた表記の表記変化のパターンを分析し、手軽に対処することを考えてオリジナルの辞書のカスタマイズという方法を採用した。実験の結果、かなり対処できていることがわかった。

なお、本稿ではオリジナルの辞書に収録済みの語に手を加えることは行わなかったが、場合によっては一部の語を削除する方がさらに性能が上がる可能性がある。また、本稿では通常の未知語のカバレッジは扱わなかったが、未知語の問題は形態素解析の性能に大きく影響する問題である。

筆者の次の課題は、本稿の成果を用いてブログテキストの「生の声」を分析することである。これについては、稿を改めて報告する予定である。

参考文献

［1］Takao, K. and Asakura, Y.: Way of Mining Travellers' Impression from Blog Pages about Using Behaviour of Kansai Airport, in *Proceedings of the 15th HKSTS International Conference*, Hong Kong, pp.703-710, 2010.

［2］鷹尾和享: 関西空港の利用行動に関する利用者の立場のブログページの簡単な収集方法, 言語処理学会第 16 回年次大会発表論文集, D2-1.pdf, 2010.

［3］森信介, 小田裕樹: 3 種類の辞書による自動単語分割の精度向上, 自然言語処理, Vol.18, No.2, pp.139-152, 2011.

［4］中田陽介, Neubig, G., 森信介, 河原達也: 点予測による形態素解析, 情報処理学会研究報告, Vol. 2010-NL-198, No.8, 2010.

［5］橋本力, 黒橋禎夫, 河原大輔, 新里圭司, 永田昌明: 構文・照応・評価情報つきブログコーパスの構築, 自然言語処理, Vol.18, No.2, pp.175-201, 2011.

［6］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, 言語処理学会第 17 回年次大会発表論文集, E4-4.pdf, 2011.

［7］齋藤邦子, 今村賢治, 松尾義博, 菊井玄一郎: 誤字脱字や伏字を許容する近似辞書照合技術, 言語処理学会第 17 回年次大会発表論文集, E5-2.pdf, 2011.

［8］MeCab: http://mecab.sourceforge.net/, 2009.

［9］UniDic: http://www.tokuteicorpus.jp/dist/, 2009.