

関西空港の利用行動に関するブログテキストの 砕けた表記とその感情の分析

鷹尾 和享¹

¹正会員 (社) システム科学研究所 (〒604-8223 京都市中京区新町通四条上ル小結棚町428 新町アイエスビル)
E-mail: takao@issr-kyoto.or.jp

近年では一般の人々によって書かれたブログが大量に出現し、書き手の生の声が含まれることが期待できる。交通行動に関しても、ブログを分析することでヒトの感情を捉えられれば、満足度の向上を図るうえで有用であると考えられる。しかし、ブログテキストには砕けた表記が多く含まれ、それらは必ずしも文法に則っていない。ところが、従来の言語処理のシステムの多くはフォーマルなテキストを中心に研究・開発されており、砕けたテキストに関しては性能が良くない。このような変化をする語は特に強調したい感情となっている可能性があり、本稿ではそのような語に着目した。本稿では、関西空港の利用行動に関するブログテキストの表記の変化を分析し、既存の形態素辞書のカスタマイズを試みた。さらに、語の意味の分類によって感情の分析を試みた。

Key Words: travel behaviour, sentiment, blog, text data, text mining, natural language processing

1. はじめに

交通行動の満足度の向上を図るためには、ヒトの心理状態を捉える必要がある。従来、交通行動に関する研究は、観測が容易な行動結果に着目されることが多かった。これに対し、ヒトの心理状態を直接的に扱う研究は必ずしも十分なされているとは言えない。

一方、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(感情)が含まれていることが期待できる。交通行動に関しても、ブログを分析することでヒトの感情を捉えられれば、交通空間の満足度の向上を図るうえで有用であると考えられる。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⁾。

テキストデータを分析するには、まず、形態素解析と呼ばれる、文章を語の最小単位(形態素)に分割する処理が手始めとなる。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、特定のデータセットである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いえない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、「すごい」が「すっごー」と変化するように、それらは必ずしも文法に則っていない。既存の形態素解析システムはこれらの箇所での解析誤りを引き起こすことが多い。さらに、このような変化をする語は特に強調したい感情となっている可能性があり、これらの解析誤りを看過することはできないと考え、本稿では、砕けた表記の語に着目した。

この問題に対処するために、正規の方法では、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。そこで、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。すなわち、既存の形態素辞書の語を流用し、表記変化したエントリを追加する方法を採った。

本稿では、砕けた表記のパターンを分析し、形態素辞書のカスタマイズによる対処方法について述べ、さらに、シソーラス辞書を用いて意味の分類をすることでそれらの語の感情を簡単に分析したので報告する。

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、UniDic⁹⁾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、砕けた表記の処理は目下盛んに研究がなされているホットなテーマであり、いくつかの成果は出ているものの、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。

一方、交通行動や都市空間に関する感情に着目した研究例としてZhang(2009)¹⁰⁾が挙げられるが、この方法は予め用意した尺度でデータ収集を行う点が本稿と異なっている。また、大塚ら(2010)¹¹⁾はアンケートの自由記述データから都市・景観イメージを分析している。

3. 表記変化の分析と形態素辞書のカスタマイズ

(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MeCab⁸⁾を用いてその形態素辞書のカスタマイズを試みる。ブログデータは、鷹尾(2011)²⁾で収集した、関西空港に関するブログ1000件のうち、利用行動を記述しているもの369件を用いる。

これ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、辞書に未登録の文字並びが未知語として出力されるが、ひらがな1文字(特に小文字)で不自然に区切れていたり、長音記号が単独で区切れていたり、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。これらの原因として、カバレッジ不足の語が多いが、語の表記変化もかなりある。長音やひらがな小文字をノイズのように見なして間投詞として辞書登録し、フォーマルな日本語の部分だけを抽出する方法も考えられる。しかし、「無理かなーと思って」のように文の途中で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、右側が本稿で追加登録する変化形である。

(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に形容詞、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したり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
楽 → ラク

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、一段動詞の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) ○○と・○○に

副詞などで「○○と」「○○に」の「○○」の部分がカタカナ化する場合がある。

例：さくっと → サクッと
ほんとに → ホントに

さらに、副詞「○○っと」の「っ」がひらがなのまま残っている場合もあるので、誤記対策として「っ」形も必要となる。逆に、オリジナルの辞書が「っ」の場合もあり、逆に「ッ」に変換させる必要のあるものもある。

例：ホロッと → ホロっと
ドバッと → ドバッと

(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、連用タ接続の活用形にみられる。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：長い → 長ーい, 長～い
長く → 長ーく, 長～く
長かつ → 長ーかつ, 長～かつ
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：長い → 長ー

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 → たあ, たー, た～
ぞ → ぞー, ぞお, 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：ずっと → ずーっと
こんな → こ～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～に
むかし → むかーし
ぶらぶら → ぶ～らぶら

(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碎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 → しいんと

(5) 形容詞語尾の脱落+引き音/促音

形容詞の終止形で語尾が脱落し、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 → 甘あ

(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, すげえ, すげー

(7) 副詞（擬態語「りと」）の促音化

副詞「〇〇りと」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

(8) です・ます

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

(9) 衍字

入力ミスなのか、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：おいしかつ → おいしいかつ

(10) 組み合わせ

以上の変化パターンが複数組み合わせられた場合もある。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～んと

(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	んかつ
連用テ接続	んかつ, んく, んくっ

(12) 対処が困難なもの

形態素辞書に登録する際には、当該形態素のコスト値が必要で、登場頻度の高い形態素ほど低コスト値となるよう設定する必要がある。本稿では、簡単のため、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。

例：ぜんぶ → ぜーんぶ

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：あつと言う間に → あつと言うまに

- ・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. 結果

(1) 規模

作成した辞書のサイズを表-2に示す。追加した形態素数が多かった要因として、カタカナ化が多数にのぼる点と、活用語の変化が多くの活用形に亘る点が挙げられる。また、本稿で述べた変化は一部の語に限られる場合もあると思われる。たとえば、「うら淋しい」のようにもともと硬い感じの語は「うら淋しっ」のように砕けた変化をすることはほとんどないと思われるが、変化する/しないの区別が困難なため、外見のパターンが該当する語全てについて一括して変化形を追加した。

(2) 性能テスト

本稿の方法の性能を見るため、テストを行った。ブログテキストからカタカナ・長音（および〜）・小文字あいうえおの並びを取り出し、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-3に示す。

「通常語ーもともと収録」「通常語ー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、「通常語ーその他の未知語」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）・一般の名詞（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。すなわち、表記変化の全22+4=26件のうち、オリジナルの辞書で対処できていたものは7件であったのに対し、本稿の方法でカスタマイズすると22件が成功している。

失敗した「表記変化一本稿で扱わなかった」ものは、(i)長音や小文字の挿入個数にきりが合わない場合（例：がんばぁあれ）、(ii)一部のみひらがな→カタカナ（例：よめはん）、(iii)外来語の表記の違い（例：エクスプレス→エキスプレス）の場合があった。

(3) 品詞別上位

次に、対象のブログテキスト全体を形態素解析し、表記変化語の数を見ることで、量的な効果をチェックした。表-4に品詞別の個数上位を示す。副詞類とは副詞と名詞-副詞可能を合算したものである。表を見ると、助詞等の変化語が多いが、副詞類や形容動詞語幹、形容詞もかなり多く、辞書追加の効果が出ていることがわかる。

表-2 辞書のサイズ

項目	形態素数
オリジナルの辞書のサイズ	392126
追加した変化形	146369

表-3 テスト結果

正否	分類	度数
○	表記変化一本稿で扱ったもの	22*
○	通常語ーもともと収録	21
○	通常語ー未知語処理で正しく解析	37
×	表記変化一本稿で扱わなかった	4
×	ハンドル名、子供の呼称等	4
×	通常語ーその他の未知語	11
×	通常語ー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

表-4 表記変化語の品詞別上位

品詞	度数	例
助詞	437	なあ、ねえ、なあ、な～、よ～、なー
副詞類	251	ワクワク、ギリギリ、ずーっと
助動詞	141	た～、たー、たあ～、ます～
感動詞	130	さあ、ああ、あー、はあ、いやあ、あ～あ
形容動詞語幹	127	ダメ、キレイ、ガラガラ、ステキ
形容詞-自立	110	イイ、でっかい、すげー、ヤバイ
代名詞	66	オイラ、コレ、ココ、ワタクシ
動詞-自立	28	ケチる、ゆ～、バてる、イける、ハマる
名詞-一般	18	ホント、ガタガタ
接続詞	6	さーて、しか～し、ケド

表-5 表記変化語の意味コード別上位（助詞等を除く）

意味コード	度数	例
195《大変》	21	ホント、すげー、ひじょ～～に
692《上機嫌》	21	ワクワク、嬉しい～、ル～ンルン
194《限度》	19	ギリギリ
635a《足取り》	18	ウロウロ、テクテク、ぶ～らぶら
132c《疎密》	17	ガラガラ、ガラ～んと
695b《無気味》	16	さすが
172a《良否》	13	マシ、イイ、イける
117a《大小》	10	でっかい、デカイ、デカイ
163《過激》	10	すげえ、ガンガン、すっごー
178a《美醜》	9	キレイ、きれ～
694b《安心》	9	イライラ、ヤキモキ
172c《良否》	8	ステキ、すーばらしい

(4) 意味の分析

次に、表記変化語の意味を調べるため、角川類語新辞典¹²⁾を用いてその意味コードを付与した。意味コード別の度数上位を表-5に示す。表を見ると、感情を込めている意味コードが上位に出ていると言える。最も多いのは195《大変》であるが、これは強調する意味であり、これらの修飾先の語を調べるとさらに感情をよりの確に分析できると思われる。また、感情を表す語自体だけでなく、その対象の「何が」を分析すればよりわかりやすくなると思われるが、これに関しては今後の課題である。

5. 結論

本稿では、ブログテキストの碎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。テストの結果、かなり対処できていることがわかった。

さらに、それらの語の意味を分析した結果、感情がこもっている意味コードが多いことが確認できた。

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、場当たりの最小限にしか行わなかったが、悪影響のある語をシステムティックに削除する方が場合によっては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。

今後の課題として、感情を表す語だけでなく、「何が」を分析することで、結果をよりわかりやすく整理できる可能性がある点が挙げられる。

参考文献

- 1) Takao, K. and Asakura, Y. : Way of mining travellers' impression from blog pages about using behaviour of Kansai Airport, *Proc. of the 15th HKSTS International Conference*, Hong Kong, pp.703-710, 2010.
- 2) 鷹尾和享 : サポートベクトルマシンを用いた関西空港の利用行動に関するブログページの収集, 土木計画学研究・講演集, Vol.44, ¥pdf¥102.pdf, 2011.
- 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.
- 10) Zhang, J. : Subjective well-being and activity-travel behavior analysis - Applying day reconstruction method to explore affective experience during travel -, *Proc. of the 14th HKSTS International Conference*, Vol.2, pp. 439-449, 2009.
- 11) 大塚裕子, 森田哲夫, 吉田朗, 小島浩, 塚田伸也 : テキストマイニングによる都市・景観イメージ分析ー水・緑環境に着目してー, 土木計画学研究・講演集, Vol.41, ¥pdf¥132.pdf, 2010.
- 12) 大野晋, 浜西正人 : 角川類語新辞典CD-ROM版, 角川書店/富士通, 1989.