マイクロブログに対する形態素・正規化情報のアノテーション

鍛治 伸裕[†]　　吉永 直樹[†]　　喜連川 優[†‡]
[†]東京大学 生産技術研究所　[‡]国立情報学研究所
{kaji,ynaga,kitsure}@.tkl.iis.u-tokyo.ac.jp

1　はじめに

近年，Twitterに代表されるマイクロブログの普及に伴い，くだけた表現を含むテキストが，自然言語処理技術の適用対象として急速に重要性を増しつつある．しかしながら，現在の形態素解析器は，そうした表現を十分高い精度で解析することが困難となっている．

くだけた表現に対して頑健な形態素解析を実現するための方法として，正規化処理と形態素解析処理を組み合わせるというアプローチが，有望な選択肢の1つとして考えられる [6, 8, 9]．また，そもそも正規化には表現の多様性を吸収するという効果があるため，正規化処理と形態素解析処理を組み合わせることによって，係り受け解析など，形態素解析に後続する解析処理の精度向上も期待することができる．

しかし，現在のところ，マイクロブログに対して形態素情報や正規化情報のアノテーションを行ったコーパスは存在していない．このため，上記のような解析器を構築しようとしても，その精度を定量的に評価することや，誤りの分析を行うことが困難になっている．

このような問題意識に基づき，我々は，マイクロブログに対して形態素および正規化情報をアノテーションしたコーパスの構築を行った．本論文では，正規化処理と形態素解析処理の組み合わせを念頭におきながら，正規化が有効であると考えられるくだけた表現の分類を行う．そして，それらのくだけた表現に対するアノテーションの仕様と実際にアノテーションされた事例の紹介を行う．さらに，アノテーションされたコーパスを使って，正規化処理が形態素解析処理に与える影響について初歩的な実証実験を行ったので，その結果についても報告する．

2　マイクロブログテキストの収集

アノテーション対象のテキストは，現在最も広く利用されているマイクロブログサービスであるTwitterへの投稿 (以下，ツイート) から収集した．具体的には，Twitter API[1]を利用して2013年12月6日に171,386の日本語ツイートを収集し，その中から無作為に選んだ1000ツイートをアノテーション対象とした．

表 1: ツイートから抽出されたテキストの例.

今日からこれで仕事します（／／／▽／／／）
アクアちゃん乗車第２号はぴーやさんとゆー笑
ありがとぉ ┌（ ┌＾ｏ＾）┐
えーくつかおうかなー
ドーナツうまそう（≧∇≦）

表 2: くだけた表現の分類と具体例.

誤字脱字	ただいm(ただいま)
	まだ絆 か(が) あるんだよなぁ
字種の変異	周囲の目が キニナリマス(気になります)
	悪ぃ(悪い) かよ
口語調表現	英語 分かんねぇ(分からない)
	取ら れとる(れている)
方言	大丈夫やで(大丈夫だよ)
	やめてくれ へん(ない) かな
略語	おわしご(しごとおわり) !!
	リプ(リプライ) もらえると

収集したツイートに対しては，アノテーションの前処理として整形処理を行った．今回は試行錯誤をしながら人手で整形処理を行ったが，コーパス構築の手続きを透明化するため，将来的には，自動整形ツールを使うことや整形処理を全く行わずにアノテーションを行うことを考えている．

我々の行った整形処理は以下の通りである．まずはじめに，ツイート中の文字を全て全角に変換し，改行や空白を取り除いた．ただし，英単語および数値表現の区切りを表す空白は残した．次に，ツイートに頻出する特殊な表現 (リツイート記号，アットメンショ

[1] https://stream.twitter.com/1.1/statuses/sample.json

ン，ハッシュタグ，URL) を取り除いた．ただし，ハッシュタグとユーザ名は，文中で 1 つの形態素として使われている場合がある．そのような場合には，ユーザ名はそのまま残し，ハッシュタグはシャープのみを削除した．最後に，非公式リツイートなどの文が途中で切れてしまっているツイートと，意味の把握が困難なツイートを削除した．以上の作業によって得られたツイートを人手で文単位に分割した結果，1831 文を得た (表 1)．

3 形態素情報と正規化情報のアノテーション

3.1 マイクロブログ上のくだけた表現

マイクロブログ上では，いわゆる整った書き言葉テキストには見られない，くだけた表現が多く使われる (表 2)．本研究では，誤字脱字，字種の変異 (大文字と小文字の違いも含む)，口語調の表現，方言，略語の 5 種類のくだけた表現に焦点を当てる．ただし，字種の変異に関しては，表 2 に見られるような，規範的な日本語表記から大きく逸脱したものに限定する．

くだけた表現を含むテキストは，くだけた表現を全く含まない，正規化されたテキストへと変換可能であると考えることができる．上記の例文においては，下線部を括弧内の表現に置き換えたものが，正規化されたテキストに相当する．このように正規化処理が施されたテキストのことを**正規テキスト**，正規化前のテキストのことを**原テキスト**と呼ぶ．

3.2 アノテーションの仕様と具体例

1 節において議論したように，マイクロブログ上のくだけた表現に頑健な形態素解析器を構築するためには，正規化処理との組み合わせ処理が重要になると考えられる．そのようなシステムに対する評価基盤を構築するため，マイクロブログに対して，原テキストと正規テキスト両方の形態素情報をアノテーションする．

我々がアノテーションを行った情報は以下の通りである．まず，原テキストに対して，形態素の表層形と品詞情報 (品詞大分類，品詞細分類，活用型，活用形) のアノテーションを行う．品詞活用体系は JUMAN 辞書 (バージョン 7.0) に従う．さらに，原テキストと正規テキストの間で表層形が異なる形態素に対しては，正規テキストにおける表層形と品詞情報もアノテーションする．以下では，正規テキストにおける表層形のことを**正規形**と呼び，単に表層形と言ったときには，原テキストにおける表層形を指すものとする．

紙面の制約上，アノテーション仕様の詳細を説明することはできないが，以下ではアノテーション例をいくつか簡単に説明する (表 3)．まず，例文 (a) は誤字の例である．表層形「か」は「が」の誤字と判断できるため，正規形を「が」とする．表層形に対する品詞情報は，正規形と同じ情報をアノテーションする．

例文 (b) は，漢字もしくは平仮名で表記するのが一般的なところを，片仮名で表記しているという，字種の変異の例である．どこまでを字種の変異として認めるのかについては，規範的な日本語から大きく逸脱しているかどうかを判断基準としているが，判断に迷う事例もみられた (例:「オンナ」)．この判断については，現在のところ作業者の主観に委ねており，今後，仕様を詳細化したい．また，字種の変異と認められた場合，正規形をどう定めるのか迷う事例もあった (例:「キタナイ」の正規形を「きたない」とするか「汚い」とするか) が，これについても今後検討する．

口語調表現としては，表記の縮約や長音化などに正規形をアノテーションする．例文 (c) では「ねえ」と「だろ」に対する正規形として「ない」と「だろう」をアノテーションする．ここで，表層形「ねえ」に対する適切な品詞情報は，JUMAN 辞書に定義されていない．そのため，活用形以外の品詞情報は正規形と共通とする．活用形は，正規形の活用形が「基本形」のため「基本形 (異表記)」とする．

一方，例文 (d) では，口語的な接尾辞「とる」を「いる」に変換するとき，正規テキストが日本語として自然なものとなるためには，直前の接尾辞「れ」も「れて」に変換するが必要がある．この接尾辞「れ」自体は，特に口語的な表現というわけではないが (c.f.,「取ら/れ/ます」)，このような場合は，接尾辞「れ」に対しても正規形をアノテーションする．

例文 (e)，(f)，(g) は方言の例である．方言は，基本的に口語調表現と同様であるが，口語的な音の変化の有無にこだわらず，例えば「チャリ」なども含めて，標準語では使われないような表現に対して正規形をアノテーションする．

次に，略語に対するアノテーション例を示す．略語はこれまで議論をしてきた表現と違って，原テキストを正規化することによって，1 形態素が複数形態素に変換される場合が多い．そうした場合には，例文 (h) のように，複数の正規形と品詞情報のアノテーションを行う．ただし，例文 (i) に示すように，略語であっても，表層形と正規形の間に 1 対 1 の対応関係が見ら

表 3: 形態素情報と正規化情報のアノテーション例．品詞情報は品詞大分類と活用形のみを記載している．

	原テキスト			正規テキスト		
	表層形	品詞大分類	活用形	表層形 (正規形)	品詞大分類	活用形
(a)	絆	名詞				
	か	助詞		が	助詞	
	ある	動詞				
(b)	周囲	名詞				
	の	助詞				
	目	名詞				
	が	助詞				
	キ	名詞		気	名詞	
	ニ	助詞		に	助詞	
	ナリ	動詞	基本連用形	なり	動詞	基本連用形
	マス	接尾辞	基本形	ます	接尾辞	基本形
(c)	泣け	動詞	未然形			
	ねぇ	接尾辞	基本形 (異表記)	ない	接尾辞	基本形
	だろ	助動詞	ダ列基本省略推量形	だろう	助動詞	ダ列基本推量形
(d)	取ら	動詞	未然形			
	れ	接尾辞	基本連用形	れて	接尾辞	タ系連用テ形
	とる	接尾辞	基本形	いる	接尾辞	基本形
(e)	大丈夫や	形容詞	ヤ形基本形	大丈夫だ	形容詞	基本形
	で	助詞		よ	助詞	
(f)	やめて	動詞	タ系連用テ形			
	くれ	接尾辞	基本連用形			
	へん	接尾辞	基本形 (異表記)	ない	接尾辞	基本形
	か	助詞				
	な	助詞				
(g)	チャリ	名詞		自転車	名詞	
	で	助詞				
	３０	名詞				
	分	接尾辞				
(h)	おわしご	名詞		しごと/おわり	名詞/名詞	
	!!	特殊				
(i)	リブ	名詞		リブライ	名詞	
	もらえる	動詞	基本形			
	と	助詞				

れる場合もある．

2 節で説明した 1831 文に対して，上記のようなアノテーションを行った結果，合計 14,323 形態素情報が付与された．そのうち，正規形が付与された形態素の数とその内訳を表に示す．ただし，1 つの形態素が複数のカテゴリに所属しうる場合があることに注意されたい．例えば「はぴば (ハッピーバースデー)」は，字種の変異であると同時に，略語でもある．複数カテゴリに所属する表現は，字種の変種と口語調表現の両方に所属しているものが最も多く，その数は 19 個であった．

正規化処理と形態素解析処理の組み合わせによる効果を見積もるため，原テキストと正規テキストに対する MeCab (バージョン 0.996)[2]の解析精度を比較した (表 5)．この結果，正規テキストの解析結果の方が，F 値が 3 ポイント以上高いという結果が得られた．これにより正規化処理と形態素解析処理を組み合わせる効果を確認することができた．

[2]http://code.google.com/p/mecab/

表 4: くだけた表現の出現回数と品詞の内訳．括弧内の数字は全形態素に対する割合を表す．

	出現回数	名詞	動詞・形容詞	その他
誤字脱字	26 (0.18%)	7	8	11
字種の変異	130 (0.91%)	60	26	44
口語調表現	549 (3.8%)	50	111	388
方言	64 (0.45%)	3	23	38
略語	117 (0.82%)	104	0	13

表 5: MeCab による原テキストと正規テキストの解析結果の比較．

	適合率		再現率	F 値
原テキスト	67,8 (11,161/16,459)		77.9 (11,161/14,323)	72.5
正規テキスト	71.7 (11,626/16,212)		80.7 (11,626/14,414)	75.9

4 関連研究

形態素情報がアノテーションされた日本語コーパスとしては，京都大学コーパス [4]，KNB コーパス [7]，BCCWJ[5] などが広く知られている．なかでも後者 2 つのコーパスには，本研究における正規形と類似する情報がアノテーションされている．KNB コーパス

においては，長音化など，口語調表現の一部に対して異表記情報がアノテーションされている．BCCWJ では，各形態素に対して代表的な表記 (語彙素) がアノテーションされている．

我々のコーパスの特徴は，マイクロブログというくだけた表現が多いドメインを対象としている点である．そのため，従来コーパスと比較して，正規化処理と形態素解析処理を組み合わせたシステムの評価に適したものとなっている．従来コーパスでも，ブログ上のくだけた表現のアノテーションは行われているが，より即時性の高いメディアであるマイクロブログの方が，くだけた表現を好む傾向が強い．例えば，KNB コーパスでは，全 66,954 形態素中，口語的な活用語の異表記が 38 回 (0.057%) 出現しているが [7]，我々のコーパスには 14,323 形態素中 129 回 (0.90%) 出現している．

KNB コーパスや BCCWJ でアノテーションされているのは，基本的には異表記に関する情報であるが，我々は，異表記に限らず，テキスト正規化に関わる情報をアノテーションしている．例えば，我々のアノテーションには「チャリ」と「自転車」や「おわしご」と「しごとおわり」のような同義表現が含まれる．さらに「取られとる」と「取られている」など，文体差に関わるような，機能語の使い分けも含まれる．

Foster ら [1] や Gimpel ら [2] は，英語ツイートに品詞タグをアノテーションしたコーパスを構築しているが，正規形のアノテーションは行われていない．一方，Han ら [3] は，英語ツイートに単語の正規形をアノテーションしているが，品詞タグの付与は行われていない．

5 おわりに

マイクロブログ上のくだけた表現に対する形態素解析処理を高精度化するためには，正規化処理との組み合わせが重要と考えられる．本研究では，そうしたシステムの評価基盤を整えるべく，マイクロブログに対して形態素情報と正規化情報のアノテーションを行った．

今後の課題としては，正規化処理と形態素解析処理を組み合わせた解析器の開発を行い，今回構築したコーパスの活用を図っていくことが挙げられる．

それと同時にアノテーションの拡充にも努めていきたい．例えば「いって (行って/言って)」など，同音異義語が存在する場合には，漢字表記を正規形としてアノテーションすることを検討している．また，マイクロブログには助詞の省略が頻繁に見られ，これが文節区切りや係り受け解析に悪影響を与えているため，省略された助詞のアノテーションにも取り組みたい．

参考文献

[1] Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. #hardtoparse: POS tagging and parsing the twitterverse. In *AAAI-11 Workshop on Analyzing Microtext*, 2011.

[2] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proc. of ACL*, pp. 42–47, 2011.

[3] Bo Han and Timothy Baldwin. Lexical normalization of short text messages: Makin sens a #twitter. In *Proc. of ACL*, pp. 368–378, 2011.

[4] Sadao Kurohashi and Makoto Nagao. Building a Japanese parsed corpus while improving the parsing system. In *Proc. of LREC*, pp. 719–724, 1998.

[5] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 2013.

[6] Ryohei Sasano, Sadao Kurohashi, and Manabu Okumura. A simple approach to unknown word processing in Japanese morphological analysis. In *Proc. of IJCNLP*, pp. 162–170, 2013.

[7] 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.

[8] 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. 言語処理学会 第 18 回年次大会, pp. 1272–1275, 2012.

[9] 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 5 回年次大会, pp. 590–512, 1999.