Twitter日本語形態素解析のためのコーパス構築

大崎彩葉　唐口翔平　大迫拓矢　佐々木俊哉　北川善彬　堺澤勇也　小町守

首都大学東京
osaki-ayaha@ed.tmu.ac.jp

1　はじめに

新聞のような整った文章に比べ，Twitter 上に存在する砕けた表現やネットスラングなどの多くの未知語を含む文章に対しては形態素解析の精度が大きく低下してしまうという問題がある．形態素解析器の精度向上のためのアプローチとして，辞書を作成する方法とコーパスを作成する方法が考えられる．しかし，砕けた表現は非常に多様であり，また SNS を始めとするネット上では日々新たな表現が生まれていることから，全てを辞書に書き尽くすことは非常に困難である．

これらのことから，SNS 用の形態素解析のための専用のコーパスを構築することで解析器の精度向上を図ることを考えた．そこで，まず日本語 Twitter テキストに対して既存の形態素解析器がどのように間違えるかを分析し，解析結果を参考にして新たに形態素情報を付与するコーパス構築にあたり，SNS の文章の形態素解析を扱うためにどのような付加情報を定義すべきかを検討し，SNS のテキストデータについて人の手でアノテーションを行い，現在の解析器では適切な結果を得られない語についてどう扱うべきか議論した．

本研究の主要な貢献は以下の 2 点である．

- Twitter 日本語形態素解析の問題点の分析と品詞体系の提案
- 形態素解析の評価や分野適応に十分な規模の Twitter 日本語形態素解析コーパスの公開

2　関連研究

Twitter などのソーシャルメディアは大量のリソースを保有しており，有益な情報を含んでいることから近年の自然言語処理の研究の対象になっている．しかし，Twitter などのウェブ文書は未知語が多く存在したり，もとの表記から崩れたりすることから，形態素の解析誤りの原因になっており，エラー分析が行われている [13]．このような，綺麗でない文書は話し言葉をそのまま文として書き下すことによるものも含まれ，今ほどソーシャルメディアが普及していない頃から，話し言葉の形態素解析として研究されてきた [8]．最近では，オノマトペのような未知語に対してはルールベースの形態素解析器 JUMAN [2] にシンプルなオノマトペの生成ルールを作ることによって対処している [6]．また，SVM などの線形分類器を用いた形態素解析器 KyTea [4, 3] にキーボード入力のログを利用して未知語の対応をしている研究もある [7]．さらに，話し言葉に存在する新しい言い回しや新語などの言語現象を分析し，後段のアプリケーションを考えた上で，品詞，読みといった適切な付加情報を考える必要がある．

日本語のコーパスは，日本語書き言葉均衡コーパス [9]（以下，BCCWJ）が広く利用されており，Yahoo! 知恵袋，Yahoo!ブログといったウェブ文書も含まれる各ジャンルに対して，単語境界と，付加情報として，品詞，活用，基本形，読みなどがアノテーションされている．その他に，大学生によって書かれたブログ 249 記事，4,186 文からなる，京都大学ブログコーパス [10]，ウェブ文書 15,000 文に対して，形態素・固有表現・構文・格関係，照応・省略関係，共参照の情報を付与した京都大学ウェブ文書リードコーパス [11] などがある．しかし，以上で挙げられるウェブ文書は比較的綺麗な文書であり，データの入手が難しいものもある．そこで，我々は，口語表現や新語などを多く含む Twitter のデータにアノテーションをし，アクセシビリティが高いデータの作成を目指した．

コーパスを作成する際は，単語分割の基準を定め，同時に品詞体系の定義をする必要がある．日本語のコーパス作成においては，BCCWJ，または，JUMAN の単語分割，品詞体系基準でコーパスを作成していることが多い．BCCWJ では，「短単位」「長単位」といった基準を定義し，コーパスの作成を行っている．JUMAN は辞書により単語分割の基準を決めており，コーパス作成の際はそこから，目的に応じて拡張を行っている [10, 1]．Kaji ら [1] のように Twitter のデータに対し

表 1: 単語分割に関する精度

	precision	recall
KyTea	86.9	90.7
アノテータ 1	97.1	92.7
アノテータ 2	98.5	93.8
アノテータ 3	97.6	93.1
アノテータ 4	91.4	95.6

て JUMAN の基準を用いてアノテーションを行っている研究もあるが，ソーシャルメディアが生み出す新たな言語現象などに焦点を絞って詳細に分析している研究は少ない．そこで，我々は Kaji らと異なる BCCWJ の短単位を基準とした KyTea [4, 3] で Twitter のデータを解析した結果を分析した．特に，[13] で言及されていなかった品詞付与に関するエラーを分析し，それらの抱える課題を整理した．さらに，Twitter に対して，複数のアノテータでの一致率を評価し，ウェブ文書コーパス作成における形態素情報のアノテーションの難易度を示した．

3 Twitter 日本語形態素解析のエラー分析

2015 年 10 月に投稿された Twitter からランダムに 100 件のツイートを抽出し，KyTea で形態素解析したデータを用意する．用意したデータを理系大学の学部生である 4 人のアノテータがそれぞれ単語分割や品詞割当が適切でない箇所を,「短単位」[1] を基準に修正した．アノテーションを開始した段階で考えていた品詞の種類は以下の 16 種類である [2]．

各人のアノテーションが完了したところで，それぞれの修正結果を比較，議論し，ツイート 100 件の形態素解析に対する gold standard data を作成した．

抽出した 100 件のツイートについて，KyTea での解析結果と作成した gold standard data で比較し，エラー分析を行い，単語分割の誤りと品詞割当の誤りごとに実際の例を集計し，適合率と再現率を求めた．結果を表 1，表 2 に示す．

3.1 単語分割の誤り

まずは，単語分割の誤りについて述べる．KyTea と複数のアノテータの適合率，再現率を比較すると，KyTea ではツイートの単語分割が困難であることがわかる．主な単語分割の誤りの種類と例を表 3 に示す．この結果は，[13] の分析結果ですでに述べられていて，本稿で分析したデータの中では特に新しい事例は見られなかった．

3.2 品詞付与の誤り

ここでは，品詞割当の誤りについて述べる．表 4 の結果から KyTea では特に，副詞・動詞・形容詞・形状詞・感動詞・接頭辞による誤りが多く見られることがわかる．しかし，これらの誤りの多くは 3.1 節で述べた単語分割に起因する品詞誤りであることがわかった．表 4 にそれぞれ例を挙げる．

またそれ以外の部分では，表記による品詞誤りが見られた．表記による品詞誤りは，ひらがなで書かれることで誘発される場合が多かった．例として，本文中では動詞の来たとして「きた」が使われているが，これを名詞の「きた」としているものが挙げられる．

4 Twitter 日本語形態素解析のためのコーパス構築

4.1 単語分割に関するエラー分析と分類

今回の集計で見られた単語分割に関するエラーは，口語的表現が含まれる文章に対するエラー（表 3：1, 8 行），ネットスラング等のマイナーな表現が含まれる文章に対するエラー（表 3：4，5，7 行），本来漢字や片仮名で表記されるものが平仮名で表記されている等の表記ゆれを含む文章に対するエラー（表 3：2, 3，6 行）のいずれか，あるいはこれらの混ざったもの（表 3：8，10 行）として分類できるケースが多かった．

4.2 品詞付与に関するエラー分析と分類

集計結果より，現在の解析器は副詞，感動詞，接続詞に関して特に精度が低い傾向が見られた．3.2 節で述べた通りその多くは単語分割に起因する誤りであり，それらは単語分割の精度の向上に伴って減ると考えられる．ここではそれ以外のケースを重点的に考える．

表 5 に例を挙げる．単語分割に起因しない品詞付与誤りで，今回の集計で多く見られたパターンとして，表記ゆれや「ー」を含む語に対するエラー，擬音語や擬態語のような名詞との区別がつき辛く，半ば造語のような語に対するエラー，ある語につく接頭辞がその前の語の接尾辞とされてしまう，複合名詞が名詞＋接尾辞，名詞＋接頭辞とされてしまうなどの接尾辞，接頭辞に関するエラーがあった．

[1] http://pj.ninjal.ac.jp/corpus_center/bccwj/morphology.html

[2] その 16 種類の内訳は副詞，助詞，動詞，名詞，空白，記号，代名詞，助動詞，形容詞，形状詞，感動詞，接尾辞，接続詞，接頭辞，連体詞，補助記号である．

表 2: 品詞付与に関する精度

	KyTea		アノテータ 1		アノテータ 2		アノテータ 3		アノテータ 4	
品詞	precison	recall	precision	recall	precision	recall	precision	recall	precision	recall
副詞	81.0	68.1	100.0	100.0	97.7	100.0	100.0	97.7	100.0	97.7
助詞	84.7	96.7	98.5	97.9	98.8	98.2	98.5	98.2	98.2	98.2
動詞	84.3	90.5	99.4	100.0	100.0	100.0	98.9	99.4	98.9	99.4
名詞	86.5	89.8	99.2	96.4	99.2	99.8	97.2	98.4	98.0	94.3
記号	0.0	0.0	75.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0
代名詞	100.0	96.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
助動詞	90.7	89.7	98.3	98.3	98.3	98.3	99.4	98.9	97.8	98.3
形容詞	87.5	84.8	100.0	100.0	100.0	100.0	100.0	100.0	97.0	100.0
形状詞	92.3	80.0	100.0	100.0	100.0	100.0	93.3	93.3	100.0	93.3
感動詞	71.4	25.0	100.0	95.0	100.0	100.0	100.0	85.0	100.0	95.0
接尾辞	85.9	96.4	98.2	98.2	100.0	98.2	100.0	98.2	94.9	98.2
接続詞	0.0	0.0	100.0	50.0	100.0	100.0	100.0	100.0	100.0	100.0
接頭辞	76.4	81.2	100.0	100.0	100.0	100.0	100.0	87.5	92.8	81.2
連体詞	85.7	100.0	100.0	100.0	100.0	100.0	85.7	100.0	100.0	100.0
補助記号	86.7	92.8	95.9	96.1	97.3	97.0	96.5	96.4	93.4	94.2
文末詞	0.0	0.0	88.8	95.2	86.1	73.8	86.6	92.8	86.3	90.4

表 3: KyTea における単語分割の誤り

誤りの種類	KyTea	gold
口語的誤り	っ/す	っす
ひらがなに起因	きた（北）	き/た（来た）
カタカナに起因	オ/ススメ	オススメ
ネットスラング	粉/み/かん	粉みかん
顔文字	(/≧/∇/≦/)	(≧∇≦)
感動詞	お/早う	お早う
固有名詞	自由/が丘	自由が丘
方言	し/たって	したって
語尾の変形	なら/な/イカ	なら/なイ/カ
品詞の誤認識	大き/い	大きい
小文字	おっ/は/よぉ	おっはよぉ

表 4: KyTea における品詞付与の誤り

品詞	KyTea	gold
副詞	も/う（助詞）	もう
動詞	きた（名詞）	き/た（来た）
形容詞	もぉい/や/や（名詞）	もぉ/いや/や
形状詞	や/だ（助詞）	や/だ
感動詞	お/早う（名詞）	お早う
接頭辞	ご/利用（接尾辞）	ご/利用

4.3 Twitter 日本語形態素解析のための品詞タグ

以上で述べたことと gold standard 作成時の議論を元に，以下のことを提案する．

口語的表現の処理について 「さ，寒い」という声を詰まらせる様子を表現したツイートに対し，一文字目の「さ」が助詞として分類されてしまう事例が見られたため，品詞分類に「フィラー」を追加した．[3]

「ごめん」や「おはよう」といった語に対し，名詞と品詞付与されている箇所があったことを受け，これらを感動詞として扱うことにした．今回 KyTea の辞書として使用した UniDic では「ごめん」は名詞とされていたが，アプリケーション的観点及び口語的表現という観点から考えると，「ごめん」は「ごめんなさい」の省略として使われることが多いため，「ごめん」単体で感動詞と見なすべきだと判断したからである．同様に，「おはよう」も「おはようございます」の形に限定せず，「おはよう」単体で感動詞と見なした．[4]

SNS 特有の表現の処理について 「なう」「だん」「わず」は Twitter 特有の語である．これらはしばしば名詞の後や文末について「〜しています」「〜にいます」「〜しました」「〜にいました」などの意味を表し，サ変動詞「する（為る）」に似た働き・意味を持つが，これらの語自体は活用形を持たないため動詞とは言えない．活用を持たず文末につくことが多いという特徴から，終助詞と同一のクラスで扱うことを考えたが，実際に形態素解析器をテキスト解析に活用する際に，これらの語が助詞として扱われるとテキストの特徴の抽出が難しくなることが予想された．そこで，終助詞を助詞から分離してこれらの語と合わせて一つの品詞分類とし，文末につくことが多い語の集まりであることから「文末詞」という名称で扱うことにした．

また，漫画のキャラクターなどが個性付けのために台詞の最後に「〜ナリ」「〜でゲソ」などの語をつけるように，「文末や文の区切れ目に付き，活用を持たず，文脈上の意味を持たない語」も「文末詞」に含めた．

KyTea では，顔文字を構成する記号列は記号本来の意味に応じて分割されてしまい，顔文字として一つの

[3] 日本語話し言葉コーパス（CSJ）では形態素情報とは別のレイヤーでアノテーションされているが，ここでは全ての形態素解析のレイヤーで扱う方針のため，このようにした．

[4] 浅原 [12] による BCCWJ の係り受け関係ラベルの定義では，フィラー・感動詞の他に，顔文字・補助記号だけでなく接続詞・非言語音・URL も係り先なしの要素となっている．

表 5: 品詞付与の誤りの分類分け

誤りの分類	語	KyTea	gold
表記ゆれ,「ー」	くださーい（「ください」の表記ゆれ）	名詞	動詞
	くっそ（「とても」の意で使われる「くそ」の表記ゆれ）	名詞	副詞
名詞との区別が難しい語	ドンドン	名詞	副詞
接尾辞，接頭辞	ご/利用（「ドンドン/ご/利用/」の形で出現）	接尾辞/名詞	接頭辞/名詞
	高速/船	名詞/接尾辞	名詞/名詞

塊と扱うことは考慮されていない．そこで，顔文字はそれを構成する記号列全体をまとめて一つの補助記号として扱った．また,「(笑)」のような形で文末につく「笑」は，文脈上の働きとしては絵文字等に近いと判断し，補助記号として扱った．

4.4 考察

「そっから」「っす」などの処理，正規形について 「そこから（そこ：代名詞＋から：助詞)」という意味で使われた「そっから」という表現が，一単語の名詞として分割されている事例や,「〜です（です：助動詞)」という意味で使われた「〜っす」という表現が「っ：語尾＋す：語尾」と分割されてしまっている事例を受け，こういった砕けた表現への対応が必要であることがわかった．こういった表現については，正規形からの派生として動的に捉えることで解決できることが知られており [5]，そのように扱うために正規形のアノテーションをする必要がある [1]．

方言の処理 「やろ（『だろ』の関西弁)」「したって(『してやって』の関西弁)」が正しく品詞がついていない事例を受け，方言への対応が必要であることがわかった．方言は予め知識がなければ人間でも正しく単語分割，品詞付与をすることは難しい．これに関してはランサーズなどのクラウドソーシングを活用することで，方言話者をリクルートすることが考えられる．

また，ひらがなで表記される語が多く，方言への対応性を高めることで他の語の表記ゆれが方言として誤認されるなど，他の解析結果に影響を与えることが危惧される．方言を含む表現の内部構造を解析するのは難しいが，方言とそれ以外が正しく分割できれば情報抽出用途ではかまわないため，方言を含む表現を左側から見た品詞と右側から見た品詞を区別するようなアノテーションを行うことにより，方言以外の箇所を正しく品詞付与できるのではないかと考える．

5 おわりに

本稿では日本語 Twitter 形態素解析のためのコーパス構築に取り組んだ，100 件のツイートに対し BCCWJ 短単位を基準として 4 人でアノテーションし，品詞付与の問題点を分析した．現在，さらに 2,000 件のデータのアノテーションを行っており，`https://github.com/tmu-nlp` にて公開予定である．また，今後は Kaji ら [1] のように日本語の正規化について取り組みたい．

参考文献

[1] Nobuhiro Kaji and Masaru Kitsuregawa. Accurate word segmentation and POS tagging for Japanese microblogs: Corpus annotation and joint modeling with lexical normalization. In *Proc. EMNLP*, pp. 99–109, 2014.

[2] Sadao Kurohashi and Daisuke Kawahara. Japanese morphological analysis system. *JUMAN version 5.1 manual*, 2005.

[3] Graham Neubig and Shinsuke Mori. Word-based partial annotation for efficient corpus construction. In *Proc. LREC*, pp. 2723–2727, 2010.

[4] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proc. ACL-HLT*, pp. 529–533, 2011.

[5] Itsumi Saito, Kugatsu Sadamitsu, Hisako Asano, and Yoshihiro Matsuo. Morphological analysis for Japanese noisy text based on character-level and word-level normalization. In *Proc. COLING*, pp. 1773–1782, 2014.

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

[7] Fumihiko Takahasi and Shinsuke Mori. Keyboard logs as natural annotations for word segmentation. In *Proc. EMNLP*, pp. 1186–1196, 2015.

[8] Kiyotaka Uchimoto, Chikashi Nobata, Atsushi Yamada, Satoshi Sekine, and Hitoshi Isahara. Morphological analysis of a large spontaneous speech corpus in Japanese. In *Proc. ACL*, pp. 479–488, 2003.

[9] 前川喜久雄. KOTONOHA『現代日本語書き言葉均衡コーパス』の開発 (< 特集 > 資料研究の現在). 日本語の研究, Vol. 4, No. 1, pp. 82–95, 2008.

[10] 橋本力, 黒橋禎夫, 河原大輔, 新里圭司, 永田昌明. 構文・照応・評判情報つきブログコーパスの構築. 言語処理学会第 15 回年次大会発表論文集, pp. 614–617, 2009.

[11] 萩行正嗣, 河原大輔, 黒橋禎夫. 多様な文書の書き始めに対する意味関係タグ付きコーパスの構築とその分析. 自然言語処理, Vol. 21, No. 2, pp. 213–247, 2014.

[12] 浅原正幸. 係り受けアノテーション基準の比較. 第 3 回コーパス日本語学ワークショップ, pp. 81–90, 2013.

[13] 鍛治伸裕, 森信介, 高橋文彦, 笹田鉄朗, 斉藤いつみ, 服部圭悟, 村脇有吾, 内海慶. 形態素解析のエラー分析. ProjectNext エラー分析ワークショップ, 2015.