Twitter 分析のための形態素解析の最適化

山田 勉
日本ユニシス株式会社　総合技術研究所
tsutomu.yamada@unisys.co.jp

1　はじめに

twitter のツイートから時系列のトレンド分析やセンチメント分析, ユーザ毎の趣味趣向分析などのツイート分析を行う際に, ツイート特有の文体や語彙, 句読点なく続く文, 顔文字・アスキーアートなどが分析の障害となっている. 自然言語解析を行う上で簡易に解析精度を高める方法は, ドメイン固有の辞書の作成とドメイン固有の未知語抽出モデルを作成することである. twitter のような流行に敏感なメディアの分析を行う場合, 新語登録の追随がトレンド分析の精度向上には不可欠である. また自然言語解析の障害となる顔文字・アスキーアートはセンチメント分析の１つの重要なパラメータであるため, できる限り正しい区切りで抽出しなければならない.

本稿では, 形態素解析用辞書の拡充と形態素解析器に未知語抽出モデルの適用を行い, 実際にツイートの分析を行った. その結果, ツイートの分析のための形態素解析の最適化の効果と課題を確認した.

2　ツイート分析

2011 年 7 月 1 日～2013 年 12 月 31 日までの 2 年半分のグローバルタイムラインから 1%の割合でランダムサンプリングした 560,337,339 ツイートを分析対象とする. ツイートには平仮名, 片仮名, 漢字のいずれかの文字が 1 文字以上含まれているものとした.

形態素解析で使用する単語辞書は NAIST-jdic Version 0.6.3 の 485,863 語から 42,553 語を削除した 459,129 語に対し, 新たに 393,368 語を追加して合計 852,497 語とした. 形態素解析プログラムは, MeCab と同等なアルゴリズムを採用し, 本稿の未知語抽出モデルを適用したプログラムを開発した.

そしてツイートを形態素解析した結果である, 形態素数 145,888,175 語, 形態素の使用回数 9,275,378,010 回を基に形態素解析の最適化の検証を行った.

3　単語辞書の拡充

他のデータソースから正確な情報を入手可能な単語については, 適宜単語を追加することにより単語辞書の鮮度を保つことができる.

事前調査より位置情報を含むツイートが多いことが確認できるため, これら地名は漏れなく辞書登録を行う必要がある. 日本国内の地名は, 日本郵便のホームページより郵便番号と地名の対応表が入手できる, 鉄道の路線名や駅名は有限個であるため, 漏れなく登録する. ただし, 組織名は, たとえば一部上場企業名などを一括して登録することが考えられるが, ツイート分析ではあまり効果がない.

新語の入手先として Wikipedia を利用する例は多いが, 本稿では,「はてなキーワード」を選択した.「はてなキーワード」は, 新語・流行語が比較的早く登録されるため, ツイート分析に適切である.「ニコニコ大百科」では, 明らかに使用されないであろう単語が多数登録されており, 単語の辞書登録の自動化が難しいと判断した.

4　未知語抽出モデル

形態素解析用の辞書に存在する語を既知語, 存在しない語を未知語とし, 入力文字列から未知語を単語として抽出する処理を未知語抽出とする. ツイート分析では, 擬態語・擬音語や顔文字・アスキーアートが形態素解析の障害となる. そのため, これらが含まれることを想定した未知語抽出モデルを作成した.

入力部分文字列に対して, 既知語があったとしても常に未知語の抽出を試みる. 未知語抽出モデルにより複数の未知語候補を生成し, 既知語と同様にラティス構造に追加する. 未知語抽出モデルは, 文字種と出現パターンによるヒューリスティックに基づきルールを作成した.

4.1 片仮名語

片仮名で表現される形態素は,外来語,擬音語,擬態語,畳語である.基本的には片仮名文字が続く範囲を未知語として切り出す.その片仮名文字種には,平仮名の濁点,半濁点,繰り返し,様々な長音記号が含まれても良いものとした.これは文の書き手はこれらの文字の意味を正しく理解せずに使用している場合があるためである.また,畳語であると判断した場合は,その繰り返しが終了する部分までを未知語とした.

品詞判定は,名詞か感動詞の判定を行った.一般的な畳語は既知語として登録済みであるとの前提で感動詞とした.感動詞の判定パラメータは,同じ文字が幾つ続くか,含まれる小書き文字の割合,WAVE_DASH が含まれるかどうか,などを使用して判定を行った.それ以外は名詞とした.

4.2 平仮名語

ツイートでは通常は漢字で記述する単語であっても,漢字を使用せずに平仮名だけを使用する人も多い.さらに平仮名だけを使用した擬音語・擬態語も使用される.平仮名の未知語抽出パラメータは,拗音や濁音の有無,“さん”,”ちゃん”などの敬称の有無,同じ文字が幾つ続くか,などで行った.品詞判定は片仮名語と同様である.また,平仮名語に関しては形態素の文字数が多いほど生起コストを大きくし,抽出される形態素の文字数があまり長くならないようにした.

4.3 顔文字(アスキーアート)

顔文字・アスキーアートでは,ギリシャ文字,キリル文字,イ文字など様々な文字種を使った新しい文字列が創作されている.そのため,顔文字として使用される文字を定義するより,顔文字として使用されない文字種とその文字種の例外文字を定義して文字列の抽出を行った.たとえば,アルファベット文字は顔文字では使用されることは少ないが,”bflovTo”は例外とした,片仮名文字は顔文字では使用されないが”ノ”は例外とするなどである.これらの例外文字は主に形態素解析で間違った結果の統計を取ることにより取得できた.

4.4 アルファベット／数字／記号

正規表現で定義できる URL,メールアドレス,日付,時刻,地番,郵便番号,アカウント,ハッシュタグは,正規表現にマッチする文字列をそれぞれ未知語として抽出した.たとえば,数字から始まる場合,日付,数値,郵便番号,記号列などを想定して抽出した文字列をそれぞれラティス構造に追加する.

5 ツイート分析結果

560,337,339 ツイートを形態素解析した結果,形態素数 145,888,175 語,形態素の使用回数 9,275,378,010 回 を基に形態素解析結果の検証を行った.まず,既知語と未知語の割合,および形態素の使用回数を表1に示す.

表 1 既知語と未知語の比率

	形態素数 (既知語と未知語の比率)	形態素使用回数 (既知語と未知語の比率)
既知語	663,053 (0.45 %)	8,280,786,471 (89.28%)
未知語	145,220,122 (99.55%)	994,591,539 (10.72%)
合計	145,888,175	9,275,378,010

既知語数の割合はわずか 0.45%であり,残りの 99.55%は未知語となっている.ツイート分析を行う場合は,未知語抽出が重要であることが確認できた.しかし,形態素の使用回数で見ると既知語が 89.28% であり,単語辞書は十分に機能していると言える.

5.1 既知語

既知語の品詞別使用比率を表2に示す.

表 2 既知語の品詞別使用比率

品詞	辞書登録数	使用形態素数 (使用率)	使用回数 (ツイート含有率)
名詞	636,712	540,211 (84.84%)	2,663,223,466 (32.16%)
動詞	174,935	99,329 **(56.78%)**	1,056,818,449 (12.76%)
助動詞	205	205 (100%)	740,318,172 (8.94%)
助詞	265	261 (98.49%)	1,817,520,236 (22.42%)
形容詞	30,408	13,582 **(44.67%)**	186,247,457 (2.25%)
副詞	3,335	3,250 (97.45%)	161,196,522 (1.95%)
接頭詞	239	236 (98.74%)	5,3614,764 (0.65%)
接続詞	191	191 (100%)	46,458,773 (0.56%)
連体詞	165	159 (96.36%)	27,797,979 (0.34%)
フィラー	18	18 (100%)	4,673,737 (0.06%)

感動詞	364	354 (97.25%)	143,616,454 (1.73%)
顔文字	998	746 (74.75%)	38,059,860 (0.46%)
英単語	4,498	4,371 (97.18%))	54,394,663 (0.66%)
記号	108	95 (87.96%)	262,108,916 (3.17%)
その他	56	45 (80.36%)	986,128,491 (11.91%)
合計	852,497	663,053 (77.78%)	8,280,786,471 (100%)

辞書に登録した形態素が実際に使用されたかどうかを示す使用率を見ると，動詞と形容詞の使用率がそれぞれ 56.78%と 44.67% であり，他の品詞と比べると使用率が低い．文語体の動詞・形容詞が使用されていないのは確認できるが，それが twitter 特有の傾向であるか，そもそも他でも使用されていない形態素なのかどうかといった検証はできていない．

次に名詞の詳細品詞別の使用比率を表 3 に示す．

表 3　名詞の詳細品詞別使用比率

名詞の 詳細品詞	辞書登録数	使用形態素数 (使用率)	使用回数 (ツイート含有率)
名詞 一般	351,029	301,847 (85.99%)	1,114,206,954 (41.84%)
固有名詞 一般	33,044	22,287 (67.45%)	107,228,485 (4.03%)
固有名詞 人名	99,990	86,989 (87.00%)	121,631,896 (4.57%)
固有名詞 組織	21,109	15,100 (71.53%)	20,205,691 (0.76%)
固有名詞 地域	106,044	89,484 (84.38%)	87,647,229 (3.29%)
サ変 副詞可能	18,278	17,493 (95.71%)	415,097,738 (15.59%)
その他	7,218	7,011 (97.13%)	797,205,473 (29.93%)
合計	636,712	540,211 (84.84%)	2,663,223,466 (100%)

辞書に拡充した形態素の中で，登録時に詳細品詞を確認できていない単語は名詞一般として登録している．固有名詞とすべき形態素も名詞一般で登録している場合があるため，名詞一般の登録数は多くなっている．

固有名詞-人名について詳細な使用率を表 4 に示す．

表 4 固有名詞-人名の使用比率

人名の 詳細品詞	辞書登録数	使用形態素数 (使用率)	使用回数 (ツイート含有率)
個人名	51,902	42,007 (87.63%)	10,819,852 (9.04%)
姓名	21,329	20,687 (96.99%)	69,246,756 (57.88%)
名前	24,293	22,765 (93.71%)	39,581,710 (33.08%)
合計	97,524	85,459 (87.635)	119,648,318 (100%)

姓名と名前はほぼ同じ割合で使用されている．個人名は，上位 500 人で 51,3%を占める．個人名の 1 位は「初音ミク」で 193,733 回，2 位は「虎徹」で 80,556 回，以下「前田敦子」，「大島優子」，「西野カナ」と続き，上位の個人名はアイドルとキャラクター名で占められていた．個人名の単語の使用率も 87.65%と高い割合ではあるが，1 単語当たりで換算すると 258 回と使用回数は少ないため，個人名の追加は単語登録の手間に比べて効果は少ない．

固有名詞-地域について詳細な使用比率を表 5 に示す．

表 5　固有名詞-地域の使用比率

地域	辞書登録数	使用形態素数 (使用率)	使用回数 (ツイート含有率)
都道府県	232	236 (97.41%)	16,962,453 (19.35%)
市区町村	2634	2,631 (99.89%)	16,005,771 (18.26%)
町域	81283	69,270 (85.22%)	29,693,528 (33.88%)
鉄道路線	1049	985 (93.90%)	1,702,769 (1.94%)
駅	9158	8,999 (98.26%)	9,052,004 (10.33%)
その他	11,688	7,373 (63.08%)	14.230,704 (16.24%)
合計	106,044	89,484 (84.38%)	87,647,229 (100%)

ツイートでは位置情報を表す語の使用率が高いことが確認できる．

固有名詞-組織は他の固有名詞と比べると単語の使用率は低い．組織名の 1 位は「JR」で 1,250,883 回，2 位は「NHK」で 372,199 回，以下「マクドナルド」，「ファミリーマート」と続き，上位 100 社で 50.25%を占める．

最後に，追加した形態素全体の使用比率を表 6 に示す．

表 6　追加した形態素の使用比率

辞書	辞書登録数	使用形態素数 (使用率)	使用回数 (ツイート含有率)
Naist- jdic	439,129	310,167 (67.56%)	6,094,193,741 (84.41%)
追加した 単語	393,368	345,165 (87.75%)	1,125,257,445 (15.59%)
合計	852,497	655,332	7,219,451,186

追加した形態素が実際に使用された率は 87.75%であった．その中で「はてなキーワード」を元にして追加した形態素は 319,537 語であり，追加した形態素の 92.57% を占める．ツイート分析用に拡充した形態素の選択は有効であったと言える．

5.2 未知語

未知語の内容別の抽出比率を表 7 に示す. 固有名詞と感動詞は形態素を構成する文字種により詳細化を行った.

表 7　未知語の抽出比率

種類	形態素数 (割合)	使用回数 (ツイート含有率)
URL	89,422,298 (61.58%)	109,902,017 (11.05%)
アカウント名	24,142,962 (16.63%)	309,090,543 (31.08%)
ハッシュタグ	1,935,853 (1.33%)	57,223,681 (5.75%)
記号,一般	6,656,251 (4.58%)	62,015,392 (6.24%)
アルファベット	4,313,801 (2.97%)	79,900,723 (8.03%)
数値	825,045 (0.57%)	56,395,099 (5.67%)
顔文字(AA)	7,001,851 (4.82%)	143,097,089 (14.39%)
固有名詞（平仮名）	72,136 (0.05%)	646,395 (0.06%)
固有名詞（片仮名）	7,596,665 (5.23%)	80,988,613 (8.14%)
固有名詞（漢字）	1,996,588 (1.37%)	43,846,285 (1.85%)
感動詞（英数字）	231,488 (0.16%)	10,508,172 (1.06%)
感動詞（平仮名）	31,688 (0.02%)	18,375,925 (1.85%)
感動詞（片仮名）	993,496 (0.68%)	22,601,605 (2.27%)
合計	145,220,122	994,591,539

ツイート分析における未知語は URL，ツイッターアカウント名が大きな割合を占めている. この 2 つだけで形態素数の 78.21%，使用回数の 42.13%となる．URL は短縮 URL が利用されるため，2 回以上使用された URL は 864 個しかなく，ほとんどが 1 回しか出現しない.

日本語ハッシュタグは通常の文として扱い, ハッシュタグから除いてある. ここでのハッシュタグは英数字を用いたハッシュタグだけである.

記号一般には, 住所の地番, 郵便番号, 電話番号, 日付, 時刻などが含まれる.

顔文字は一部を辞書に登録している．1 語当たりのツイートでの使用回数を比べると, 辞書に登録している顔文字は 51,019 回と使用回数が多いが，未知語の顔文字は 20 回と少ない．一方，形態素数で比較すると，辞書登録している 746 語に対して，未知語の顔文字は 7,001,851 語であり, 様々なバリエーションの顔文字が未知語抽出されていることが分かる.

固有名詞（平仮名）は主に「さん」「ちゃん」などの敬称を含む未知語が抽出された．ただし，小書き文字から始まる未知語が 34,195 語あり，これは未知語抽出の間違いであると考えられる.

固有名詞（漢字）はおそらく中国語であろう未知語がほとんどであった. 中国語のツイートが混入していたためであるが, 使用されている文字種だけでそれが日本語なのか中国語なのか判断することは困難である. 若干ではあるが人名の名前が抽出できていた.

感動詞（平仮名）は, 使用回数の上位の未知語は 1 文字, もしくは 2 文字であり，それだけで使用回数の 69.85%を占める．これがすべて感動詞であるとは考えにくく，単語の一部が未知語として抽出されたものである.

それ以外の未知語に関しては大きな問題はなかった.

6　おわりに

本稿では, ツイート分析のための形態素解析の最適化を行い, 実際にツイート分析を行った結果から効果の確認ができたことを述べた. 新語の登録は,「はてなキーワード」を利用すると新語の使用率が高い. 未知語抽出モデルは顔文字やアスキーアートを含んだツイート分析に効果がある．この手法は同じような文体・語彙が使用されるブログや, アスキーアートが多用されるメルマガの分析にも適用できる.

未知語抽出モデルにおいて, 平仮名から構成される未知語の抽出に課題が残った. 平仮名の未知語の抽出を間違えると，その後の構文解析での影響が大きい．そのため，平仮名の未知語抽出モデルの改良を今後の課題としたい.

参考文献

[1] 萩原正人,関根聡. 翻字と言語モデル投影を用いた高精度な単語分割. 言語処理学会 第 19 回年次大会

[2] 乾孝司, “不自然言語処理”, 情報処理 Vol.53 No.3 Mar. 2012 p202-p203

[3] 笹野遼平, 鍛治伸裕, “新しい語・崩れた表記の処理”, 情報処理 Vol.53 No.3 Mar. 2012 p211-p216