Twitterを用いた時制を表す特徴語の自動収集に関する研究

赤﨑優介　　森田和宏　　泓田正雄　　青江順一
徳島大学大学院　先端技術科学教育部

1. はじめに

Web 上には電子掲示板やブログなど，人々が容易に情報を発信できるツールが存在する．特に Twitter に代表されるマイクロブログでは，個々人の考えや行動などの情報をリアルタイムに発信することができる．Twitter とは，ツイートと呼ばれる 140 文字以内の短文をパソコンや携帯端末から投稿できる情報サービスであり，リアルタイム性や波及性に優れているという特徴がある．そのため，話題抽出[1]や情報伝播の分析[2]など様々な研究に利用されている．また，Twitter はジオタグを付与した投稿が可能であり，経度と緯度で表された自身の現在位置をツイートと共に発信することができる．このジオタグ機能を用いることで，人々の行動や，行動を伴う発言の解析をすることができる．例えば，“これから大阪に向けて出発する”という発言の後，実際に大阪からの発言があれば，この文には未来を表す表現が含まれているということがわかる．時制を表す表現は，ユーザが考えている未来の予定や過去の出来事などの情報を得る際の手掛かりとなる．

そこで，本研究では Twitter とジオタグ機能を利用して時制を指し示す特徴的な語を収集すること，また，特徴的な語を用いて文の時制を判定することを目的とする．

2. 関連研究

Twitter のジオタグ機能を用いた研究として，酒巻ら[3]はユーザの行動パターン調査に関する研究をおこなっている．任意のユーザが特定の場所でおこなうツイートを解析することで，その場所がユーザにとってどのような意味を持つかを推定するというものである．推定の対象はジオタグを付与した投稿を日常的におこなっているユーザであるため，ジオタグ機能を利用していないユーザに対しては推定をおこなえないという問題がある．また，山田ら[4]は，ツイートと行動の関係を確率モデルで表現し，ユーザの未来における行動をベイズ推定によって予測する手法を提案している．この手法は，過去に同じ行動をおこなった複数のユーザの記録を元にし，同じ行動をおこなったユーザが未来におこなう行動を予測するものである．そのため，過去に同じ行動をおこなったユーザの情報をある程度取得していなければ予測できないという問題がある．

そこで，未来や現在を指し示す特徴的な語を元に文書を分類することで，ジオタグを普段用いていないユーザに対しても行動の情報を得ることが可能だと考えられる．

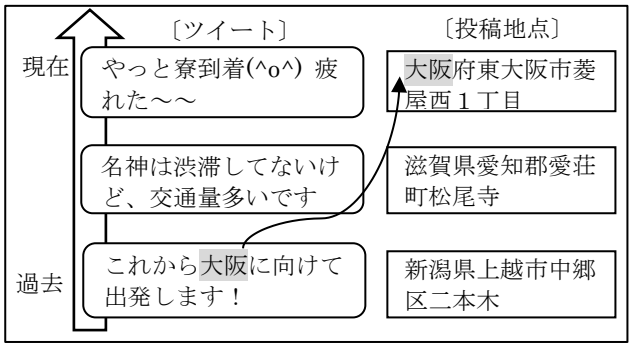

図１：ツイートと投稿地点の比較

3. Twitterを用いた文の分類

時制を特徴づける語の取得には，文書群から学習をおこなうことが必要となる．本研究では，はじめに，特定の地名を指すツイートと，そのツイートの後に発信されたツイートの投稿地点を比較することで，未来の予定や現在・過去の出来事を指し示す文書群を取得する(図 1)．以降，本稿では未来を表す文を＜未来＞，現在・過去を表す文を＜現在＞とする．

3.1 文の収集

TwitterAPI を用いてユーザの投稿群を取得し，ジオタグが付与されたツイートを時系列順に得る．その際，リツイート内容を含むツイートや，位置情報サービスから投稿されたツイートなどは除去する．次に，@ユーザ名や URL など，ノイズとなる文字列の除去をおこなう．最後に，Yahoo!リバースジオコーダ API[5]を用いて各ツイートがおこなわれた投稿地点の住所を，ジオタグによる経度・緯度から得る．

3.2 文の分類

ジオタグが付与された各ツイートに対し，以下の処理をおこないツイートに含まれる文を＜未来＞，＜現在＞とそれ以外に分類する．文中に含まれる地名の判定には，晃昇ら[6]が提案した地域連想語辞書を用いる．地域連想語とは，地名や特産品など特定の地域を連想できる語のことである．

Step1.　地域連想語を使用して，ツイート本文の都道府県名を判定し，地域連想語を含む文を取得する．

Step2. Step1 で取得した都道府県名が，ジオタグから取得した投稿地点の都道府県名と一致する場合，文を＜現在＞に分類し，一致しない場合 Step3 へ．

Step3. 投稿時から 5 日後までのツイートのジオタグから取得した投稿地点を調べ，Step1 で取得した都道府県名と一致する都道府県名があれば，文を＜未来＞に分類する．

4. 特徴語の抽出

第 3 章で述べた手法により分類した文から，時制を表す特徴語の抽出をおこなう．本手法では特徴語として単語を収集するほか，単語に続く助詞や助動詞などの品詞および品詞の活用に着目するため，文節の収集もおこなう．

4.1 単語と文節の取得

学習文に対し形態素解析をおこない，各文節に対して以下の処理をおこない単語と文節を収集する．

- 体言を含む文節

文節と，文節に含まれる体言を取得する．

例）明日から → 「明日から」，「明日」

- 用言を含む文節

文末表記に重点を置くため，文の最後に位置する文節のみ取得する．その際，形容詞と形容動詞，及び動詞の語幹・活用部分を以下のように統一して収集する

“[形]” + “[(活用名)]” + 残りの品詞

“[動]” + “[(活用名)]” + 残りの品詞

例）帰ります → [動][連用]ます

ただし，状態動詞は他の動詞（変化動詞など）と品詞構成が同じ場合でも時制に違いが見られるため，一部の状態動詞(図 2)は活用が終止形の場合，以下のように統一する．

“[状]” + “[終止]” + 残りの品詞

また，「ね」などの終助詞は省略し，「て」と「で」のような意味が同じ接続助詞は統一して収集する．

4.2 出現率と偏りを考慮した特徴語の抽出

特徴語の抽出には，各分野における単語の出現率とカイ二乗値による偏りを考慮する方法[6]を用いる．はじめに，分野 k における単語 i の出現率 Y_i を次式より求める．

$$Y_i = \frac{w_i}{\sum_i w_{ik}}$$

w_{ik}：分野 k における単語 i の出現頻度

次に，カイ二乗値を求める．カイ二乗値とは，期待される出現頻度と実際の観測値の差を示す統計量であり，値が大きいほど特定の分野に偏って出現しているということがわかる．以下に式を示す．

$$\chi_{ik}^2 = \frac{(w_{ik} - m_{ik})|w_{ik} - m_{ik}|}{m_{ik}}$$

$$m_{ik} = \frac{\sum_{k=1}^{K} w_{ik}}{\sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}} \sum_{i=1}^{N} w_{ik}$$

w_{ik}：分野 k における単語 i の出現頻度

m_{ik}：分野 k における単語 i の理論頻度

N: 学習文書内の異なり総文節数　K: 分野数

以上の式より求められる出現率とカイ二乗値に対し，本手法ではそれぞれ閾値 α，β を設定し，閾値以上の単語および文節を特徴語として抽出する．また，分野 k における特徴語 i のスコア X_{ik} を以下の式より定義する．

$$X_{ik} = \frac{Y_{ik}}{\sum_k Y_{ik}}$$

5. 特徴語を用いた文の判定手法

4 章で抽出した特徴語を用いて，以下の手順により新規文に対して分野の判定をおこなう．

Step1. 新規文に含まれる文節と単語を 4.1 節と同様の方法で取得する．

Step2. 取得した語の出現率を以下の式より求める．

$$Z_i = \frac{z_i}{\sum_i z_{ik}}$$

z_i：新規文書に含まれる語 i の出現頻度

Step3. 取得した語が 4.2 節で抽出した特徴語である場合，以下の式より各分野のスコアを加算する．

特徴語iの分野kに対するスコア $= X_{ik} \times Z_i$

Step4. スコアが最も高い分野を，新規文の分野とする．特徴語が含まれていない場合は分類不可とする．

6. 実験

6.1 実験設定

はじめに，ジオタグを付与したツイートをおこなっているユーザ 587 人の投稿群を収集し，3 章で述べた方法を用いて＜未来＞，＜現在＞の文を収集した．その際，ツイッター特有の言葉である「なう」を含む文は除外した．「なう」はツイッターにおいて頻繁に用いられる言葉であり，大きな偏りが発生するためである．そして，＜未来＞，＜現在＞の文を各 5,900 文用いて，特徴語を取得した．取得した特徴語の一部を表 1 と表 2 に示す．

次に，新規文の分野判定実験をおこなった．実験には，

思う，おもう，居る，分かる，わかる，違う，悩む，出来る，できる，感じる，見える，使える，おる，…

図 2：状態動詞の一部

表 1：＜未来＞の偏りが大きい語の一部

体言を含む		用言を含む	
文節	χ^2_{ik}	文節	χ^2_{ik}
これから	43.23	[動][連用]ます	47.93
今から	39.04	[動][終止]	14.96
今	9.22	[動][未然]う	4.37
明日は	45.42	[動][未然]うと	1.44
高速バス	7.53	[動][連用]てきます	10.61
新幹線で	5.90	[動][連用]てくる	1.75

表 2：＜現在＞の偏りが大きい語の一部

体言を含む		用言を含む	
語	χ^2_{ik}	語	χ^2_{ik}
到着	95.26	[動][連用]た	50.81
やっと	6.33	[動][連用]てきた	19.23
雨が	4.02	[動][連用]ました	17.92
終了	6.28	[動][未然]れました	2.08
地下鉄	14.87	[動][連用]てる	6.69
通過	7.44	[形][終止]	11.54

表 3：閾値 α による特徴語数

α	0.10	0.15	0.20	0.25	0.30	0.35	0.40
語数	102	71	55	44	39	38	31

閾値 β を 2.0 に固定し閾値 α を 0.1 から 0.4 まで 0.05 刻みで設定して取得した特徴語を用いた．それぞれの閾値により取得した特徴語の数を表 3 に示す．判定対象には人手により＜未来＞，＜現在＞とそれ以外に分類した新規の文 500 文を用い，＜未来＞と＜現在＞の文における適合率と再現率を求めた．結果を図 3 と図 4 に示す．

6.2 考察

投稿文に含まれる地名と，投稿地点の違いを基準に学習文を収集したため，特徴語抽出では「今から」や「到着」などの語以外に，「高速バス」や「新幹線で」など移動手段に関する語が多く抽出された．

分野判定実験では，特徴語が一つしか含まれていない文の場合，その特徴語の時制が文の時制となる．しかし，特徴語以外の語により文全体の時制が変わる場合，誤分類さ

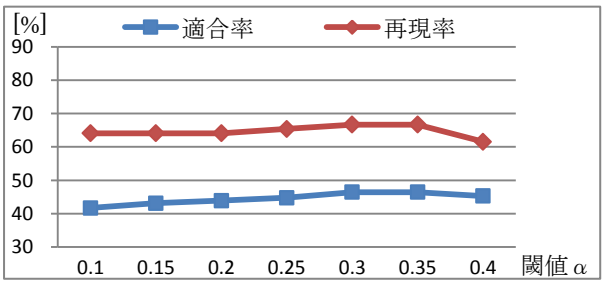

図 3：＜未来＞の精度

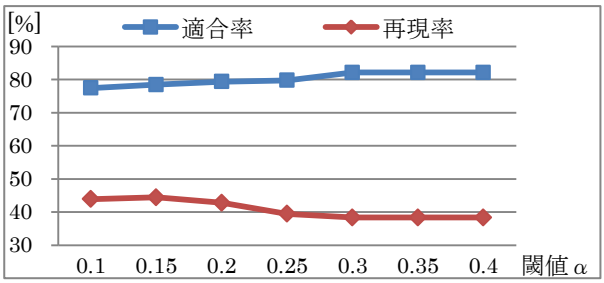

図 4：＜現在＞の精度

れるという問題がある．例えば，“空いてる”という文節は＜現在＞の特徴語“[動][連用]てる”であるが，「明後日は空いてる」という文の時制は＜未来＞である．この場合，“明後日”という単語が＜未来＞の特徴語として抽出されていなければ，＜現在＞に誤分類される．改善策としては，漢字や平仮名による表記違いの単語を統一することが考えられる．また，今回の実験では，用いることができた学習文が少なく，取得できた特徴語数が十分ではなかった．そのため，学習に用いるデータをさらに増やし実験をおこなう必要がある．

7. まとめと今後の課題

本稿では，Twitter の投稿時間と投稿地点を利用した文の分類手法と，時制を指し示す特徴語の自動収集手法，また，特徴語を用いた新規文の分類手法について提案した．今後は，問題点の改善をおこなう．

参考文献

[1] 中本聖也，北野光一，寺口敏生，田中成典，西江将男：”マイクロブログからの地域の話題抽出に関する研究”，情報処理学会第 73 回全国大会，pp.783-785，2011

[2] 風間陽一，今田美幸，柏木啓一郎：”Twitter の情報伝播ネットワークの分析”，人工知能学会第 24 回全国大会，2010

[3] 酒巻智宏，岩井将行，瀬崎薫：“マイクロブログのジオタグを用いたユーザの行動パターンの調査に関する研究”，情報処理学会第 73 回全国大会，pp.787-789，2011

[4] 山田和貴，斉藤裕樹：“マイクロブログサービスの位置情報タグと発言コンテキスト解析を用いた行動推定システムの設計”情報処理学会研究報告,Vol.2010-DBS-151, No.21,pp.1-6,2010.

[5] Yahoo!デベロッパーネットワーク-Yahoo!リバースジオコーダ API，http://developer.yahoo.co.jp/webapi/map/

[6] 晃昇祥恵，森田和宏，泓田正雄，青江順一：“地域連想語辞書の構築に関する研究”言語処理学会第 18 回年次大会 2012