Vol.2010-DBS-151 No.22
2010/11/13

ミニブログにおける注目語抽出手法の提案と注目語を用いたメディア間での話題追跡

加藤慶一[†1]　秋岡明香[†1]
村岡洋一[†1]　山名早人[†1]

Twitter に代表されるミニブログは新たなメディアとして注目を集めており，そこでの発言の解析や，テレビや新聞等の他のメディアとミニブログとの相関関係の解析に対する需要が高まっている．しかし，ミニブログにおける発言は，特定の作品や商品に関する言及を多く含み，これらの多くは複合語であるため，そもそも形態素解析を正しく行なうことが難しい．そこで，ミニブログにおける出現頻度が急上昇した自立語，特に名詞に注目し，複合語で構成される固有名詞（注目語）を取得する手法を提案する．提案手法により，ミニブログにおける形態素解析の精度向上が期待でき，ミニブログや他のメディアでの関連ある話題をより正確に追跡することが可能となる．

Extraction of Distinctive Phrases from Mini Blog Entries and Application for Topic Tracking across the Media

Norikazu Kato,[†1] Sayaka Akioka,[†1]
Yoichi Muraoka[†1] and Hayato Yamana [†1]

A mini blog service, including Twitter, is one of emerging media of note. Across-the-board analysis in posted blogs, and descriptions in related media, such as TV, newspapers, and other media, is indispensable for social analysis. Posts in mini blogs, however, often include names of particular movies, novels, and products, and many of which are compounders. A compounder is often divided into several words by word processors, and difficult to extract as one solid word. Here, if a hot compounder is extracted as it is supposed to be, the quality of morphological analysis is improved to contribute to better topic tracking in many descriptions in media. Therefore, in this paper, we propose a methodology to extract distinctive phrases from mini blog posts. The methodology picks up burgeoning keywords included in posts in limted time, and tries to form meaningful compounders.

1. はじめに

Twitter に代表されるミニブログは新たなメディアとして注目を集めており，そこでの発言の解析や，テレビや新聞等の他のメディアとミニブログとの相関関係の解析に対する需要が高まっている．たとえば，あるテレビの報道がきっかけとなり，ミニブログユーザの間でどのような議論が展開されているか，あるいは，ある製品の宣伝をテレビで行なった結果，どのようなミニブログユーザ層がどういった反応を示すか，などといった調査は，社会分析を行なう上で非常に重要である．

しかし，ミニブログにおける発言は，特定の作品や商品に関する言及を多く含み，これらの多くは複合語であるため，そもそも形態素解析を正しく行なうことが難しい．ミニブログやテレビ報道を対象とした社会分析において，新出の複合語や固有名詞を正しく抽出することは，不可欠である．そこで本稿では，ミニブログにおける出現頻度が急上昇したフレーズ，特に自立語や名詞に注目し，複合語で構成される固有名詞（新語）を取得する手法を提案する．提案手法により，ミニブログにおける形態素解析の精度向上が期待でき，ミニブログや他のメディアでの関連ある話題をより正確に追跡することが可能となる．また，Twitter およびテレビ番組の文字放送のデータを用いた評価実験により，提案手法の有効性を示すと同時に，提案手法により抽出した新語を学習させた辞書を用いて 2010 年 7 月の参議院議員選挙に関する話題追跡を行なった結果を報告する．

なお，本稿の構成は以下の通りである．第 2 章で，形態素解析手法，ミニブログを用いた社会分析，文書の話題追跡手法のそれぞれについて，関連研究を紹介する．第 3 章で，提案する新語抽出手法の詳細を述べ，Twitter を用いた評価実験を行なう．第 4 章では，提案手法を用いて抽出した注目語を学習させた辞書を用いて，Twitter およびテレビ番組の文字情報を形態素解析することで，Twitter やテレビ番組におけるトピックの解析を行ない，第 5 章でまとめる．

2. 関連研究

2.1 形態素解析

語の抽出や複合語に関する研究は既に多くあるが，ここでは代表的な研究について述べる．

†1 早稲田大学
Waseda University

湯本らは，単名詞を対象とした専門用語抽出の手法を提案している[1]．提案手法では，連接する単名詞にスコア付けを行なうことで，複合名詞の抽出を行なう．したがって，湯本らの手法では名詞同士の連接のみを扱うことになり，助詞等の名詞以外の品詞に属する語を対象とすることができない．

福島らは，形態素解析を行ない，活用語尾の性質を利用することで,「ググる」,「ウザい」などのカタカナ用言の抽出を行なっている[2]．用語の抽出にはカタカナ部分の出現回数を利用しており，その後に続くひらがなの活用部分のパターンを用いて，用言の獲得を行っている．本研究ではカタカナだけでなく，漢字なども含めた語の出現頻度を利用しており，カタカナ表現以外の語も取得することが可能である．

鍛冶らは，大規模時系列ウェブアーカイブから，新造語を分析，抽出する手法を提案している[3]．Twitter も大規模時系列ウェブアーカイブのひとつと考えることができ，アプローチとしては本稿の提案手法と同様である．しかしながら，鍛冶らの手法では新造語，すなわち動詞や名詞，形容詞などの自立語を取得することを目指しており，複数の語で構成された固有名詞を取得している本論文と目的，取得方法が異なる．

2.2 ミニブログを用いた社会分析

Cha ら[4]，Kwak ら[5]，Banerjee ら[6]，The Web Ecology Project[7]，Weng ら[8]，Java ら[9]，Zhao ら[10]，Krishnamurthy ら[11] など，Twitter の利用者やツイートの内容に関する研究は増えている．いずれの研究も，Twitter 利用者数の変遷，Twitter 利用者同士のフォロー/フォロワー関係に関する調査，利用者の知名度やリツイートされるツイートの数の関係，影響力のある利用者についての考察，地域によるトピックの変遷の比較などを詳細に調査して報告しており，資料的価値も高い．

しかし一方で，こうした従来研究は Twitter 内部での利用者同士の関係や，トピックの変遷の調査にとどまっている．現実には，ツイートの内容は Twitter 内部のみならず，インターネット上のその他の情報や，テレビなどの各種メディアに関連したり，影響を受けたりすることが多い．また，Twitter は今や主要なメディアの一部となっており，他のメディアから入手した情報を元に，利用者間での様々な議論に発展することもしばしばである．こうした背景を踏まえて，本稿では，Twitter でのツイートのみならず，テレビでの報道も含めて解析を行なう．これにより，情報がどこから発生し，どのような経路を辿って広まったか，また情報が伝搬する過程で，どのような議論が展開されたかを観察することが可能となる．

2.3 話題追跡

文書群における話題追跡手法や，文書内に記載された内容ごとの分類手法については，従来より多くの研究がなされているが，未だ汎用的かつ決定的な手法は確立されていない．近年では，たとえば Mei らが probabilistic latent semantic analysis (PLSA) モデルを拡張した確率モデルを提案している[12]．また，Steyvers らは学術論文に注目し，筆者と論文との関係についてベイズ推定を用いたモデルを導入することで論文の分類を行なう，教師なし学習アルゴリズムを提案している[13]．Wang らは，時系列に沿った語の共起に注目し，文書をトピック毎に分類する手法を提案している[14]．

話題追跡手法や文書のトピック分類を自動化する手法は，膨大な情報を効率的に処理する上で不可欠である．しかし，ミニブログなどにおいては，話題追跡以前の問題として，そもそも形態素解析を正しく行なうことができない，筆者とトピックの間に強い相関関係が存在するとは限らない，話題が短時間で頻繁に変遷する，などの問題があり，こうした問題点を解決することが，ミニブログにおける話題追跡手法確立への第一歩となる．

3. ミニブログの形態素解析

3.1 ミニブログの形態素解析における問題点

Twitter などのミニブログに投稿されるツイートは，映画やテレビ番組の話題や，特定の商品に関する言及などを多く含む．特に，映画の公開日や新商品の発売日には，映画のタイトルや商品名が数多くツイートに含まれる．映画のタイトルである「沈まぬ太陽」や「カールじいさんの空飛ぶ家」，書籍名である「きょうの猫村さん」や「坂の上の雲」などは，実際に過去のツイートに数多く出現した代表的な例である．

投稿内容を解析するためには，形態素解析が必要となる．しかし，こうした映画のタイトルや書籍名などは複合語で構成されており，正しく形態素解析を行なうことが難しい．例えば,「カールじいさんの空飛ぶ家」を形態素解析すると,「/カール/じいさん/の/空/飛ぶ/家」と品詞分解される.「カール」,「じいさん」,「空」,「飛ぶ」,「家」の共起確率を見ることで,これらの語が一緒に使われる事が多いと判断することは可能ではあるが,「カールじいさんの空飛ぶ家」という固有のフレーズとしてとらえることは難しい．

そこで本稿では，Twitter におけるツイートにおける出現頻度が急上昇した自立語，特に名詞に注目し，複合語で構成される固有名詞を取得する手法を提案する．注目を集める新規の固有名詞を正しく切り出すことで，Twitter，テレビ番組，新聞などの複数メディアで言及されている場面を正しく特定することが可能となる．また，ある特定の商品や作品に言及している場面を的確に切り出すことで，その商品や作品に関するコメントのコンテキスト解析の精度が向上することも期待できる．

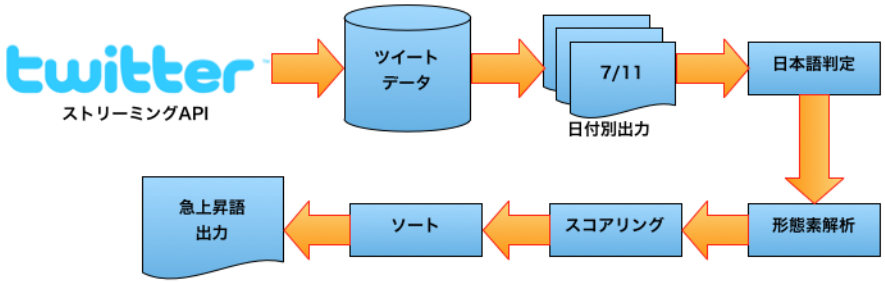

図 1 提案手法における急上昇語の抽出手順

3.2 新規固有名詞を抽出する提案手法

提案手法では，最初に Twitter のツイートにおいて，出現頻度が急上昇している「急上昇語」を取り出し，この急上昇語と近隣の語を解析することで新規の固有名詞を取得する．以下で，それぞれの手順の詳細について述べる．

3.2.1 急上昇語の抽出

急上昇語を抽出する手順について，概要を図 1 に示し，詳細を以下に述べる．

まず，Twitter から取得したツイートは，通常の形態素解析を行なうことで単語単位に分割される．

次に，各単語について，出現頻度の増減を 1 日単位で比較する．この比較を行なうために，それぞれの単語について各日付のスコアを付す．特定の単語についてのある日のスコア S_{w_d} は，以下のように計算する．

まず始めにその名詞がその日の日本語ツイート数に対する割合を求める．R_{w_d} は 1 日における語の割合，w_d は 1 日における語の出現回数，J_d は 1 日の日本語ツイートの回数を表す．最後に，このままではきわめて小さい数が出てしまうので，すべての割合に 10000 を乗する．式は以下のようになる．

$$R_{w_d} = \frac{w_d}{J_d} \times 10000 \tag{1}$$

続いて，その日の割合の平均値と標準偏差を求め，それぞれ μ_d，σ_d とする．

以上の数値を利用して，ある語に対するその日のスコアを S_{w_d} とし，以下のように求める．

$$S_{w_d} = \frac{R_{w_d} - \mu_d}{\sigma_d} \times 10 + 50 \tag{2}$$

さらに，各日で求めた S_{w_d} について，対象期間単位での出現頻度上昇の勾配を調べるために，対象期間の S_{w_d} の最大値を S_{w_dmax}，最小値を S_{w_dmin} とし，次式のようにある語についての対象期間におけるスコア S_w を求める．

$$S_w = \frac{S_{w_dmax}}{S_{w_dmin}} \tag{3}$$

このスコア S_w を降順でソートし，以下で述べる固有名詞の取得を行う．

3.2.2 固有名詞の取得

3.3.2 にて抽出した急上昇語について，複合語を構成する単語の一部であるかどうかを解析し，対象の単語が複合語の一部であった場合には，前後の適切な単語と合わせて複合語とする．たとえば,「カール」という急上昇語に対して,「今日新宿でカールじいさんと空飛ぶ家を見る」というツイートがある場合，1)「カール」が「今日新宿で」と合わせて複合語を構成するのか，2)「カール」が「じいさんと空飛ぶ家を見る」と合わせて複合語を構成するのか，あるいは，3)「カール」は単名詞であるのか，を調べることになる．具体的な手順は，以下となる．

(1) 急上昇語を含むツイートを抽出する．
(2) 急上昇語より前の文章と後の文章に分割する．
(3) 分割後のそれぞれの文章について形態素解析を行ない，形態素解析により抽出された各単語について,「急上昇語から n 個目」の形態素としてラベル付けを行なう．
(4) 急上昇語に近い形態素から，各ラベルでもっとも多い形態素が，急上昇語を含むツイートに対して含まれている割合を求める．
(5) その割合が閾値以下となるまで探索を行なう．(今回は予備実験に基づき，閾値を 0.6 とした．)

3.3 評 価 実 験

提案する新規固有名詞の抽出手法を，Twitter のツイートを用いて評価する．利用するデータは，Twitter が提供する Streaming API [*1] である．Streaming API には数種類あり，本実験ではユーザ ID のみで，無料で利用できる Sample Streaming を利用している．Sample Streaming は全ユーザの最新の投稿からランダムで選ばれた一部のデータをストリーミング形式で提供されているデータである．この他に同じ条件で利用できる API として，自分のフォローしているタイムラインの情報を得る ChirpUserStreams が存在するが，

⋆1 http://dev.twitter.com/pages/streaming_api

データに偏りが生じる可能性が高いため，今回は使用していない．

この Sample Streaming は各ツイートごとに JSON 形式でまとめた状態を送信しており，現在以下のデータを保存している．

- ユーザ ID
- ツイート ID
- スクリーンネーム
- ユーザ名
- クライアント名
- ツイート本文
- 投稿日時
- 返信先のツイート ID

以上の情報以外にも図 2 のように JSON データにユーザが入力した紹介文やアイコン画像の URL，フォロー，フォロワー数，位置情報などが含まれている．

今回の実験で利用したデータは 2010 年 7 月 10 日から 7 月 12 日の間に取得した 4,752,287 件のツイートである．本実験では日本語を対象としているが，現在，sample streaming を用いて取得できるツイートは，1 分間で約 700 件である．さらに，これらは全ての言語が含まれたツイートであるため，対象となる日本語のツイートは，11%程度の 526,998 件であった．

3.3.1 抽出された急上昇語

2010 年 7 月 10 日から 7 月 12 日の間には，参議院議員選挙や FIFA ワールドカップ決勝といった大きなイベントがあった．そこで，提案手法により抽出した急上昇語上位 20 件と実験期間にあった，選挙，ワールドカップに関連した語 20 件のを表 1 に，上位 20 件以降で選挙，ワールドカップに関連した急上昇語を 2 に示す．

上位 20 件の急上昇語はほぼ選挙，またはワールドカップ関連の語が取得できていることがわかる．一部「雨」や，ハッシュタグを表す「#」など，関連しない語もとれている．しかしながら，選挙，ワールドカップ関連のツイートにはハッシュタグがつけられているものが多く，日本全体が注目するようなイベントが無い日より，ハッシュタグが比較的多く出現したことが原因だと考えられる．急上昇語 15 位の「タコ」も FIFA ワールドカップ関連の語であり，これはドイツにある水族館のタコ，パウル君が予想した試合の勝敗をすべて的中させたため，話題となったため，出現頻度が急上昇したと考えられる．なお「パウル」は急上昇語 59 位に位置する．


```
{"text":"Writing papers", "coordinates":null,
"in_reply_to_screen_name":null,
"source":"web", "created_at":"Sun May 09 22:51:26 +0000 2010",
"contributors":null, "truncated":false, "in_reply_to_status_id":null,
"favorited":false, "place":null,
"user":{
    "notifications":null, "favourites_count":1434,
    "profile_image_url":"http://a3.twimg.com/profile_images/Ahiru.jpg",
    "profile_text_color":"3E4415", "time_zone":"Tokyo",
    "screen_name":"s_wool","statuses_count":22224,
    "profile_link_color":"D02B55", "description":"I love Apple!",
    "profile_background_image_url":"http://s.twimg.com/themes/bg.gif",
    "created_at":"Sat May 19 15:59:52 +0000 2007",
    "contributors_enabled":false,
    "profile_sidebar_fill_color":"99CC33", "lang":"en",
    "profile_background_tile":true, "location":"Japan",
    "following":null, "profile_sidebar_border_color":"D3D2CF",
    "followers_count":742, "protected":false,
    "verified":false, "geo_enabled":false,
    "name":"Loriann Daniel ", "friends_count":630,
    "id":6162582, "utc_offset":32400,
    "profile_background_color":"EDECE9",
    "url":"http://steelroom.blogspot.com/",
},
"id":13692226457, "geo":null, "in_reply_to_user_id":null}
```


図 2 取得できる JSON フォーマットの例

表 1 抽出した急上昇語上位 20 件

順位	急上昇語	順位	急上昇語
1	選挙	11	党
2	投票	12	雨
3	スペイン	13	優勝
4	24	14	政治
5	2010	15	タコ
6	オランダ	16	票
7	wc	17	速報
8	当選	18	谷
9	サッカー	19	#
10	決勝	20	当確

表 2 上位 20 件以下の急上昇語（選挙・W 杯関連）

順位	急上昇語	順位	急上昇語
22	テレ	45	日本
25	民主	46	杯
27	池上	48	議員
30	亮子	49	W
32	ワールドカップ	54	落選
34	イニエスタ	58	開票
39	みんな	59	パウル
41	自民	64	民主党
42	ドイツ	68	自民党
43	国民	72	千葉

上位 20 件以降に抽出された 22 位の「テレ」,「池上」の 2 語は，11 日のテレビ東京で放送された選挙特番「池上彰の選挙スペシャル」に関連する語である．観測期間である 7 月 10 日から 12 日の間で，それぞれの「池上」という語を含むツイートの数は，10 日が約 10 件に対して，11 日が約 600 件，12 日が約 200 件と急上昇していることがわかる．さらに，Twitter で話題になっただけではなく，実際に視聴率も選挙特番の中で NHK，日本テレビに次いで 3 位となっている．[⋆1]

3.3.2 抽出された固有名詞

続いて表 3 に，提案手法により取得した固有名詞の例を示す．

これは 3.3.2 で述べた急上昇語上位 200 件の語に対して，3.2.2 で示した手法で抽出した結果である．上位 200 件の語の中には同一の固有名詞を構成する語が含まれている．たとえば議員名である「蓮舫」の場合，mecab にて形態素解析をすると「蓮」と「舫」に分割される．今回の急上昇語上位 200 件の中に,「蓮」と「舫」の両方の語が入っていたため，取得した固有名詞には「蓮舫」が重複して出力された．表 3 には重複している語は省略したが,「蓮舫」以外に「龍馬」,「白戸次郎」などが重複していた．

まず取得に成功した固有名詞について述べる．

抽出された「テレ東」,「池上彰」は 3.3.2 で述べた選挙特番の結果からも明らかである．抽出語の中の「http://twitpic.com/24」は，保存しているツイートから検索し，目視した結果，その池上彰氏の選挙特番のシーンが twitpic にアップロードされた URL の一部であることがわかった．

選挙関連以外に，実験期間に行われたイベントに関連する固有名詞も取得できており,「白戸次郎」,「@SHIRATOJIRO」，がそれである.「白戸次郎」および「@SHIRATOJIRO」はソフトバンクが参院選にあわせて行ったプロモーションに関する語である．

このように人名，およびテレビ番組などが取得できている例があるが，URL，ハッシュタグの一部，一般複合名詞も同じように多く取得されている．

取得された URL，ハッシュタグの中でも，Twitter 連携サービスである「ツイッター診断メーカー」[⋆2] の URL が多く取得されていることがわかる．このサービスはユーザが自分でオリジナルの診断内容を作成し，他のユーザが作成された診断で遊び，その結果を Twitter に反映させることができる．ツイッター診断メーカーは 8 月 12 日時点で 37,000 以上の診断が作成され，もっとも利用回数の多い診断は 70 万回以上利用されている．診断結果には，URL が追加された内容が反映されるので，URL に診断結果の定型文が付加された形で抽出された結果も得られた.「なるほど SUNDAY じゃねーの」も Twitter を利用した「なるほど 4 時じゃねーの」[⋆3] というサービスによるもので，これはある時間（午前 4 時と日曜日）になると登録したユーザアカウントで定型文が自動投稿されるサービスである．定型文は午前 4 時に「なるほど 4 時じゃねーの」と投稿され，日曜日には「なるほど SUNDAY じゃねーの」と投稿される．

今回対象期間の 7 月 12 日が日曜日であった．そのため,「なるほど SUNDAY じゃねーの」は 12 日に急上昇したと判断され，今回の実験で取得することができた．

もう一方の「なるほど 4 時じゃねーの」という内容は，毎日ほぼ同数投稿されているため，における日付間のスコア S_{w_d} の比がほぼ一定のため，急上昇語には含まれなかった．

「(#KODAKUMINETliveathttp://ustre.am/kpqK)」も USTREAM で行われたライブ中継のハッシュタグおよび URL である．この語が含まれるツイートが 7 月 10 日が 40 件であるのに対し，11 日に約 500 件と急上昇しているが，このライブ中継が 7 月 7 日から 11 日までの 5 日間連続して放送されたもので，11 日に急上昇した要因をツイートの内容から得ることはできなかった.「皆既日食」も同じように 7 月 11 日にイースター島で観測されたものがライブ中継されており，話題になっていたことがわかった．

一般複合名詞には参議院選挙関連で急上昇した「出口調査」や「消費税」などの一般的な

⋆1 http://www.itmedia.co.jp/news/articles/1007/12/news090.html

⋆2 http://shindanmaker.com/

⋆3 http://4ji.ssig33.com/

複合名詞も今回の結果に含まれていた．

以上に述べたものも Twitter 上での流行を表す表現ではあるが，固有名詞取得の観点からはこれらの語を除外していく必要があると考えられる．

3.4 評　　価

3.3.2 にて提案手法を用いて取得できた固有名詞を形態素解析辞書に追加し，急上昇語の再取得を試みた．3.3.2 と同様に上位 200 件のデータを利用して取得した結果を形態素解析辞書に追加した場合と追加前の辞書を利用した場合を比較を行う．具体的には提案手法により，それぞれの上位 200 件の急上昇語にどの程度固有名詞を含んでいるのか，人手による確認を行った．固有名詞は複合語で構成されていないもの（国名など）も含む．

確認の結果，提案手法を用いて作成した辞書を利用して取得できた固有名詞は 33 件，変換前の辞書を利用して抽出した固有名詞は 26 件という結果となり，約 27%の増加が見られた．

取得した固有名詞が増加した理由は，提案手法により新たに獲得した固有名詞が急上昇語に含まれていると同時に，こうした固有名詞は従来では分割され，人名や番組名として抽出されなかったからである．

この実験により，提案手法を用いることで話題性の高いフレーズを抽出し，より正確な形態素解析を実現できることが示せた．しかしながら「カンブリア宮殿」のような番組名の場合,「カンブリア」と省略して記述する場合もあり，提案手法を用いた場合，急上昇語のスコアが下がってしまう問題なども見られるため，こうしたケースへ今後は対応して予定である．

4. 多メディアにおける話題追跡への応用

ここでは，提案手法によりミニブログから抽出した注目語を，他のメディアにおける解析に応用する際の効果について検証する．

3.4 における評価と同様に，提案手法を用いて得られた語を追加した形態素解析辞書と，従来の形態素解析辞書のそれぞれを用いて，ミニブログから取得した注目語語を利用し，3.4 で Twitter データを収集したのと同期間に放送された文字放送について，解析を行った．具体的には，提案手法を用いた場合と従来の形態素解析辞書を用いた場合のそれぞれについて，対象期間の文字放送から抽出した注目語の数を比較した．

実験の結果．テレビの文字放送から抽出された注目語は，提案手法が 12,348 件，従来手法が 11,532 件であり，従来の形態素解析辞書を用いた場合に対して，提案手法を用いた場合には 7 %の向上がみられた．また，注目語が文字情報に含まれている語のほとんどは参議院議員選挙，FIFA ワールドカップに関係する語であることがわかった．

表 4　語句の増加に影響する差が出た名詞

急上昇語	件数
消費税	408
参院選	96
蓮舫	68
谷亮子	56
出口調査	38
トーレス	10

表 4 に提案手法と従来手法で差がついた原因となる語を示す．この語はすべて表 3 にもある語であり，提案手法によって形態素解析辞書に追加された語により，より多くの語を取得できたことがわかる．

特に,「蓮舫」,「谷亮子」,「出口調査」の語は従来手法ではそれぞれ「蓮 」と「 舫」,「谷」と「亮子」,「出口」と「調査」に分割されてしまい，今回の選挙に関連しない語となってしまう．一方で提案手法では独立した語句として取得できたためより多くの語の取得に貢献した．

この結果より，提案手法は時系列上で多大な注目を集めている話題を象徴するフレーズを有効に抽出しており，また抽出した語は他のメディア解析にも有効であることが分かった．また，同時期の Twitter データおよびテレビの文字放送を解析したことで，Twitter を用いた注目語の抽出はテレビ報道における注目語の抽出にも有意義であることを示すことができた．さらに，Twitter で注目を集めている話題とテレビで注目を集めている話題との間には一定の相関があることを確認することができた．

5. ま と め

近年では Twitter などのミニブログが注目を集めており，そこでの発言の解析や，テレビや新聞等の他のメディアとミニブログとの相関関係を解析し，社会現象を読み解くことへの需要が高まっている．しかし，ミニブログによる発言は，特定の作品や商品に関する言及や，そのコミュニティでのみ通じる造語を多く含み，従来の形態素解析辞書を用いて，発言を正しく解析することは難しい．

そこで，本稿では，ミニブログにおける出現頻度が急激に上昇した自立語，特に名詞に

表 3 提案手法により取得した固有名詞

取得された語	内容	取得された語	内容
http://twitpic.com/24	URL	#_ki	ハッシュタグ
2010wc	ハッシュタグ	出口調査	一般名詞（選挙関連）
#worldcup	ハッシュタグ	#senkyonow	ハッシュタグ（選挙関連）
テレ東	テレビ局（選挙関連）	から』です。http://shindanmaker.com/32010#_ki	URL
RT@	Twitter 特有表現	http://twitpic.com/249	URL
谷亮子	人名（選挙関連）	(#KODAKUMINETliveathttp://ustre.am/kpqK)	ハッシュタグ
#esp	ハッシュタグ	消費税	一般名詞（選挙関連）
選挙特番	一般名詞（選挙関連）	蓮舫	人名（選挙関連）
#ned	ハッシュタグ	さんが当選した暁には『	Twitter 特有表現
龍馬	人名	池上彰	人名（選挙関連）
#senkyo	ハッシュタグ（選挙関連）	カンブリア宮殿	番組名
白戸次郎	キャラクター名	します。http://shindanmaker.com/31644	Twitter 特有表現
http://nico.ms/lv	URL	千葉景子	人名（選挙関連）
。http://shindanmaker.com/	URL	投票用紙	一般名詞（選挙関連）
トーレス	人名	皆既日食	一般名詞
@SHIRATOJIRO	キャラクター名	#tvtokyo	ハッシュタグ
参院選	一般名詞（選挙関連）	#softbank	ハッシュタグ
龍馬伝	テレビ番組名	なるほど SUNDAY じゃねーの	Twitter 特有表現

注目し，複合語で構成される固有名詞や注目フレーズを抽出する手法を提案した．Twitter の Sample Streaming を用いた実験では，従来よりも 27%多く固有名詞を抽出することに成功し，これらの多くは実験期間中に開催されたワールドカップや参議院議員選挙に関連する語であることが確認できた．また，Twitter から抽出した注目語を含む辞書を用いて，同期間のテレビの文字放送を解析した結果，従来手法と比較して 7%多い注目語を抽出することができた．また，ここでも，抽出した注目語の多くは参議院銀選挙やワールドカップに関連する語であった．こうした実験から，Twitter というひとつのメディアから抽出した注目語は，テレビという他のメディアで注目されている語を抽出するにも有効であり，さらに各メディアで注目される話題には相関関係があることを再確認することができた．

今後は，注目語の抽出精度向上，注目語抽出時の閾値に関する考察と裏付けを行なって行くと同時に，本稿で得られた成果を多メディアにおけるトピック追跡に応用していく予定である．

謝辞 本研究の一部は，文部科学省 次世代 IT 基盤構築のための研究開発「Web 社会分析基盤ソフトウェアの研究開発」“多メディア Web 解析基盤の構築及び社会分析ソフトウェアの開発”によるものである．また，テレビの文字情報取得について，国立情報学研究所の佐藤真一教授の協力していただいた．ここに記して謝意を示す．

参 考 文 献

1) 湯本紘彰, 森辰則, 中川裕志, “出現頻度と連接頻度に基づく専門用語抽出”, 情報処理学会自然言語処理研究会報告 2001(86), pp.111-118, 2001 年 9 月.
2) 福島健一, 鍛冶伸裕, 喜連川優, “機械学習を用いたカタカナ用言の獲得”, 言語処理学会第 13 回年次大会発表論文集, 2007 年.
3) 鍛冶伸裕, 宇野良子, 喜連川優, “言語学研究の支援を目的とした大規模時系列ウェブアーカイブからの新造語のマイニング”, 第 1 回データ工学と情報マネジメントに関するフォーラム (DEIM2009) 予稿集, 2009.
4) Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi, “Measuring User Influence in Twitter: The Million Follower Fallacy”, Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM2010), 2010.
5) Haewoon Kwak, Changhyun Kee, Hosung Park, and Sue Moon, “What is Twitter, a Social Network or a News Media?”, Proceedings of the 19th International Conference on World Wide Web (WWW2010), 2010.
6) Nilanjan Banerjee, Dipanjan Chakraborty, Koustuv Dasgupta, Sumit Mittal,

Vol.2010-DBS-151 No.22
2010/11/13

Seema Nagar, Angshu Rai, and Sameer Madan, “User Interests in Social Media Sites: An Exploration with Micro-blogs”, Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM2009), 2009.

7) Alex Leavitt, Even Burchard, David Fisher, and Sam Gilbert, “The Influentials: New Approaches for Analyzing Influence on Twitter”, the Web Ecology Project, http://www.webechologyproject.ort/2009/09/analyzing-influence-on-twitter/, 2009.

8) Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He, “TwitterRank: Finding Topic-sensitive Influential Twitterers”, Proceedings of the 3rd International Conference on Web Search and Data Mining (WSDM2010), 2010.

9) Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng, “Why We Twitter: Understanding Microblogging Usage and Communities”, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, 2007.

10) Dejin Zhao and Mary Beth Rosson, “How and Why People Twitter: The Role that Micro-blogging Plays in Informal Communication at Work”, Proceedings of the ACM 2009 International Conference on Supporting group work, 2009.

11) Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt, “A Few Chirps About Twitter”, Proceedings of the first workshop on Online social networks (WOSN’08), 2008.

12) Qiaozhu Mei, and ChengXiang Zhai, “A Mixture Model for Contextual Text Mining”, Proc. The Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD2006), 2006.

13) Mark Steyvers, Padhraic Smyth, and Thomas Griffiths, “Probabilistic Author-Topic Models for Information Discovery”, Proc. The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD2004), 2004.

14) Xuerui Wang, and Andrew McCallum, “Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends”, Proc. The Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD2006), 2006.