社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－DBS－143 (35)
2007／7／3

ブログ文書集合を用いた省略語抽出手法の検討

関口裕一郎† 佐藤 吉秀† 川島 晴美† 奥田 英範†

† 日本電信電話株式会社 NTT サイバーソリューション研究所
〒 239-0847 神奈川県横須賀市光の丘 1-1
E-mail: †{sekiguch.yuichiro,sato.yoshihide,kawashima.harumi,okuda.hidenori}@lab.ntt.co.jp

あらまし ブログの急速な普及により，人々の生の体験や経験の情報がネットワーク上で多く発信されるようになり，それを用いたブログにおける話題の抽出等のマーケティング分析のニーズが増えてきている．しかしブログ記事は口語的な表現で記述されるため，分析時に商品名等の重要な固有表現が省略して表記されることによる分析精度の低下が問題となっている．本論文では，固有表現の正式表記の一部の文字を用いて作られる省略語を自動抽出することを目指し，ブログ文書での語句の使われ方を見ることにより省略語としての確からしさを算出手法を提案し，実際のブログ文書に適応した際の有効性について論じる．

キーワード 省略語抽出，ブログ分析，データマイニング

Clipped word extraction using blog documents.

Yuichiro SEKIGUCHI†, Yoshihide SATOU†, Harumi KAWASHIMA†, and Hidenori OKUDA†

† NTT Cyber Solutions Laboratories, Nippon Telegraph and Telephone Corporation
Hikari-no-Oka 1-1, Yokosuka-shi, Kanagawa, 239-0847 Japan
E-mail: †{sekiguch.yuichiro,sato.yoshihide,kawashima.harumi,okuda.hidenori}@lab.ntt.co.jp

Abstract Many people write their experiments and impressions in their weblogs, and these articles have a much effect on buying behavior in web shopping. Thus, there are needs for mining topics in weblog articles for marketing purpose. In such mining processes, the proper noun is very important, though, many proper nouns are written in clipped word in weblogs. We describe a method to extract clipped words of the given proper noun using weblog articles that contains the original proper noun or candidates of clipped words. And evaluate the effectiveness using large weblog corpus.

Key words clipped word extraction, weblog, data mining

1. はじめに

近年，ブログの浸透により多数の一般の人々が感想や体験を記事としてネットワーク上に発信するようになると同時に，ウェブベースでのショッピングの一般層への普及がおきている．そのような中で，ブログで発信される人々の体験情報は，商品の購入判断に用いるロコミ情報として幅広く利用されるようになり，ブログ上での評判の善し悪しが人々の購買行動に大きく影響を及ぼすようになってきている．このため商品を提供する企業において，日々更新されるブログ記事集合において特定の商品や分野に関する話題がどのように推移しているかを分析する，ブログ記事集合のマーケティング分析技術へのニーズが高まってきている．

商品に関連する話題をマイニングするにあたっては，分析対象となる商品名や競合商品の名称，販売者等の組織名，広告やキャンペーンに起用されているタレント名といった固有名詞の抽出が非常に重要である．従来はあらかじめ人手による固有名詞辞書の整備を始め，ウェブ文書を用いた専門用語の自動学習[4]や，各種固有表現抽出アルゴリズムにより，抽出精度の向上が行われてきた．しかしブログの記事では，多くの場合口語的でくだけた表現が多く用いられる為，文の構造を手がかりとした手法の適用は難しい．また，辞書に登録されている正式名称だけではなく，正式名称から派生した略称や愛称といった異表記語も同一の事柄を表す語句として一般的に用いられる．従来の正式名称を収集した固有名詞辞書ではこれらの異表記語への対処が不十分であるため，ある製品名を対象としてブログ記事のマーケティング分析を行う際などに，略称や愛称などの正式名称以外の表記で記述しているブログ記事が分析対象から外れてしまうという問題点があった．

本論文では，上に述べたような略称や愛称といった固有名詞の異表記語を，ブログ文書集合中から自動的に抽出する手法について取り扱う．特にブログ中で多く出現する，固有名詞の一

部の文字を用いることによって生成される省略語に注目し，正式名称が与えられた際にその省略語表記を自動的に抽出する手法を提案する．

以降，2章ではブログで多く用いられる表記ゆれのタイプや先行手法について説明し，3章で本論文で用いる手法について解説する．4章で提案手法の評価手法及び評価結果について述べた上で考察を行い，5章でまとめと今後の課題について述べる．

2. 固有名詞の表記ゆれ

2.1 ブログ記事中における固有名詞の異表記パターン

まず最初に，ブログ文書集合において，固有名詞の異表記がどのように出現しているのかを概観する．

2006年8月1日に書かれたブログ記事集合のうちの2000記事を対象として，各記事において同一の事柄を表す固有名詞が異なる表記で出現するパターンを人手で抽出した．集計対象とした2000記事中の367記事中に固有名詞と異表記語の組が存在し，全部で480組が存在した．その結果から，ブログで多く見られる固有名詞の異表記語を，その生成パターンに基づいて大まかに分類すると，以下の4種類に分類が出来る．

（1） 省略語：固有名詞を構成する一部の文字を抜き出すことによって作られる語句．例えば，『厚生労働省』と『厚労省』や，『中日ドラゴンズ』と『ドラゴンズ』等がある．

（2） 読み換え語：固有名詞の漢字部分をその読みとなるひらがな・カタカナに置き換えたり，英語部分をその読みとなるカタカナに置き換えることにより作られる語句．例えば，『東京都』と『とうきょうと』や，『PlayStation』と『プレイステーション』等がある．

（3） カタカナ異表記語：外国語由来の固有名詞における，読みをカタカナで表記する場合における表記ゆれによって作られる語句．例えば『ヴェネツィア』と『ヴェネチア』等．

（4） その他異表記語：上記のパターン以外の変更によって作られた語句．愛称などの場合が多い．例えば，『松任谷由実』と『ユーミン』や，『ペ・ヨンジュン』と『ヨン様』等．

上記の分類ごとの出現数を集計した結果が，表1となる．また，今回『PlayStation』と『プレステ』の組のように，読み換え語でありかつ省略語である場合は，その両方にカウントしたために，各項目の合計は480組よりも多くなっている．

表1に示されるように，最も多く現れたパターンは省略語の関係となっている組で，57.3%の275組であった．このうち，『安倍晋三』と『安倍』といったフルネームと姓のみ，名のみといった組や，『東京都』と『東京』といった地名から都道府県部分や市町村部分を除いた組のような，双方が一般的な固有名詞として扱われるパターンが44組含まれる．

以上のことから，ブログ記事上での固有名詞の異表記の多くは省略語のパターンであるといえる．また2番目に多い読み換え語については，固有名詞辞書作成時にその読み情報も作成しておくことで容易に対処可能なパターンである為，本論文では省略語の自動抽出にフォーカスして論じることとする．

2.2 関連研究

元の正式名称から省略語を抽出する手法として，正式名称の構成語句を形態素解析し，その構成要素の関連性を判断した上で，一部の構成要素の頭文字を取得する形で省略語を抽出する技術が存在する．[3] 例えば『厚生労働省』という語句であれば，「厚生」「労働」「省」という3つの単語から構成されると判断できる為，それぞれから1文字を抜き出して『厚労省』といった省略語を作成することが出来る．このような手法においては，あらかじめ定められたルールに当てはまらない省略語を抽出できない問題点がある．またブログで扱われるような新語に多く見られる，元となる固有名詞が形態素解析されないような造語などの場合に，精度が低くなるという問題点があった．

表1 ブログ記事中における異表記語のパターンことの分布

タイプ	出現数	割合 [%]
読み換え	159	33.1
省略	275	57.3
カタカナ異表記	34	7.1
その他異表記	55	11.5

また他の異表記語のパターンについての従来研究として，カタカナ語句の表記揺れを扱った研究が数多く行われている．「ツィ」と「チ」のような頻繁に起こるカタカナ異表記の変換ルールをあらかじめ作成することによりカタカナ異表記を求める手法が提案されている．[1] また，それを発展した手法として，表記違いを表記ペナルティとして数値評価した上で，ウェブ上の文書を活用した各表記の出現する文脈の類似度合いを考慮することにより，同義となる語句を抽出する手法が提案されている．[2]

3. 提案手法

省略語は元の固有名詞と同義であるため，省略語が用いられている文書は固有名詞が用いられている文書と類似していると考えられる．一方，ブログ記事集合を用いることにより，語句の使用例はふんだんに取得が可能である．

以上のことから，固有名詞の正式名称から一般に使われる省略語を取得する手法として，

（1） 正式名称を元に省略語となる可能性のある省略語候補語句を全て作成する

（2） 各候補語句を含む文書集合と，元となる固有名詞を含む文書集合をブログ記事データベースから求め，それら2つの文書集合の類似度合いを数値化する．この類似度合いの数値が高いほど，各候補語句の省略語らしさが高いと判定する

という2ステップから構成される省略語抽出手法を提案する．

3.1 省略語候補語句の作成

このステップでは単純に元となる固有名詞の文字列から，文字の順序を入れ替えずに任意の数の文字を抜き出すことにより作成できる全ての文字列を省略語候補として取り出す．

例えば『厚生労働省』という固有名詞からは，『厚』『生』等の1文字からなる候補語句が5語句，『厚生』『厚労』等の2文字からなる候補語句が10語句，『厚生省』『厚労省』等の3文字からなる候補語句が10語句，『厚労働省』等の4文字からなる候補語句が5語句作成され，全体で30語句の省略語候補語句

が作成されることとなる．

このような手法で省略語の候補語句を作成すると，あらかじめ想定されないような候補語句も作成できる一方で，極めて冗長な数の語句が作成される問題がある．これに対処する為に，各候補語句を含む文書数をブログ記事データベースから検索することにより取得し，その値が極端に低い候補語句については，一般に用いられない表記として候補から除く処理を行う．

3.2 略語スコアの算出手法

各省略語候補語句を含むブログ記事集合と，元となる固有名詞を含むブログ記事集合の，2つの文書集合が内容的に類似している場合，省略語候補語句が省略語である可能性が高いと判別することとする．そのため，2つの文書集合間の類似度合いを数値化したものを，略語らしさを表す略語スコアとすることとする．

ある省略語候補語句 w_{cl} が元となる固有名詞 w_{ne} に対して省略語となる確からしさを表す省略語スコア $S(w_{ne}, w_{cl})$ は式1によって求められる．

$$S(w_{ne}, w_{cl}) = average\left\{\sum_{d_i \in C_{ne}} \sum_{d_j \in C_{cl}} Sim(d_i, d_j)\right\} \quad (1)$$

この時，元となる固有名詞を含む文書集合を C_{ne} とし，スコアの算出対象となる省略語候補語句を含む文書集合を C_{cl} とする．また関数 $Sim(d_i, d_j)$ は，文書 d_i と d_j の類似度を返す関数であり，各文書の構成語句ではられる語句ベクトルの類似度をコサイン類似度で数値化した値を返すこととする．

またこの際に，多くの省略語候補語句は元の固有名詞の一部分となる．(例えば『最高裁判所』と『最高裁』など) このような場合には，C_{cl} に C_{ne} が含まれる形になる為，両方の集合に含まれる記事が存在することとなる．この場合重複する記事の数が省略語スコア $S(w_{ne}, w_{cl})$ に大きく影響を与えてしまう為，省略語句が元の固有名詞の連続した一部分になる場合には，元の固有名詞を含み省略語候補を含まない文書集合を C_{ne} とすることとする．

3.3 実装手法

以上で述べた省略語取得手法を，ウェブからのクローリングによって作成したブログ記事データベースを用いて実装した．

使用したブログデータベースには，2006年7月から9月の間に投稿されたブログ記事約3500万件が蓄積されており，任意の語句による全文検索が可能となっている．全文検索のインデックスは，形態素解析を利用した転置インデックスによるものとなっている．

また略語スコアを求める際には，各語句を含む検索結果上位30記事の文書集合を用いて算出することとした．また検索結果が50記事以下となる省略語候補に関しては，一般に用いられない表現と解釈してスコアの算出対象外とすることとした．

4. 評価実験

提案手法の有効さを，スポーツチーム，企業名，中央省庁名などの固有名詞20語についての省略語の正解を用意して評価を行った．使用した固有名詞と，各固有名詞についての正解となる省略語の個数が表2となる．一番省略語の正解が多かった固有名詞は『東北楽天ゴールデンイーグルス』の7語(『楽天』,『イーグルス』,『楽天イーグルス』,『楽天ゴールデンイーグルス』,『ゴールデンイーグルス』『東北楽天イーグルス』,『東北楽天』)だった．

これらの固有名詞から作成される全ての省略語候補語句に対して，提案手法による略語スコアの算出を行った．また含まれる記事が50件以下となる省略語候補語句に対しては，スコア0とした．

表2 評価に使用した固有名詞

固有名詞	省略語数
東京ヤクルトスワローズ	5
横浜ベイスターズ	5
中日ドラゴンズ	2
北海道日本ハムファイターズ	6
福岡ソフトバンクホークス	5
東北楽天ゴールデンイーグルス	7
浦和レッドダイヤモンズ	3
川崎フロンターレ	3
ガンバ大阪	2
清水エスパルス	2
ジュビロ磐田	2
鹿島アントラーズ	2
キヤノン株式会社	1
トヨタ自動車	2
松下電器産業	2
第一生命保険	1
厚生労働省	1
農林水産省	1
経済産業省	1
国土交通省	1

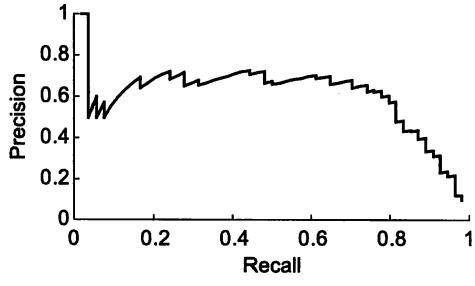

図1 適合率と再現率の関係

表3 『中日ドラゴンズ』に対する略語スコア上位6語

略語	スコア	DF
ドラゴンズ	0.240721	12449
中日	0.164542	62505
日ドラゴンズ	0.110058	59
ドラズ	0.058762	61
ゴンズ	0.057527	82
ゴズ	0.05516	67

表 4 『松下電器産業』に対する略語スコア上位 6 語

略語	スコア	DF
松下電器	0.183896	2075
松下	0.163502	19361
電器	0.156678	9715
産業	0.066035	56059
松下電産	0.046077	69
下電産	0.046042	70

表 5 『経済産業省』に対する略語スコア上位 6 語

略語	スコア	DF
産業	0.051695	54341
経済省	0.048644	62
経済	0.044584	158396
経産	0.043804	1305
産業省	0.042928	133
経産省	0.042591	1097

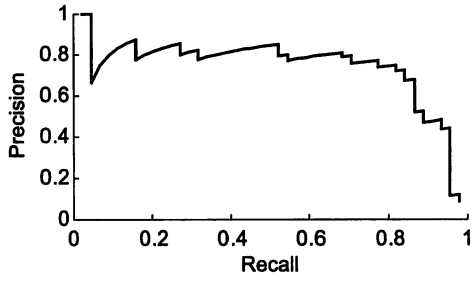

図 2 スポーツチームのみでの適合率と再現率の関係

また，算出された略語スコアに対して閾値を設定し，閾値以上の値となった場合に，その候補語句が省略語であると判定する事とした．この際に，省略語と判定された語句と，あらかじめ作成された正解である省略語と一致度合いを見ることにより，適合率と再現率を算出した．閾値を 0 まで変化させた際の適合率と再現率の関係をグラフに纏めたのが図 1 である．

再現率 74.1%の時に適合率 65.6%となり，F 値が 69.6 で最大となる．このときの略語スコアの値は 0.077 となる．低再現率の部分で一時的に適合率が落ち込む範囲があるが，全体として再現率 75%までの範囲で適合率 65%前後を達成している．

また『中日ドラゴンズ』『松下電器産業』『経済産業省』について省略語スコアを算出した場合の上位 6 つの語句と，それぞれの語句の DF 値を表 3〜5 に示す．

『中日ドラゴンズ』『松下電器産業』については，それぞれ正解である『中日』と『ドラゴンズ』,『松下電器』と『松下』，といった語句が上位に来ていることが分かる．また，略語スコア 0.15 前後で正解と不正解の分かれ目になっていることが分かる．一方『経済産業省』ついては，正解の『経済産業省』は 6 番目のスコアとなっており，また 0.052 のスコアが最高と全体として低い略語スコアしか算出されていない．

元の固有名詞の分野ごとの傾向として，スポーツチームは比較的精度よく抽出できるが，中央省庁名は上手く抽出できないという結果が出た．実際にスポーツチーム名のみに絞って，再現率と適合率の関係を求めた結果が図 2 である．こちらは再現率 70%前後までの範囲で適合率 80%前後を維持しており，全体についてと比べ精度よく抽出できていることが分かる．略語スコアの閾値を 0.073 にした際に F 値が 77.9 で最大となり，そのときの適合率が 72.5%で，再現率が 84.1%となる．

これは各固有名詞がブログ中でどのような記事として扱われるかに強く依存していると考えられる．スポーツチーム名とその略称は多くの場合，試合の感想の記事で扱われるため，それらの記事で使われる語句は似通う傾向が強く，結果スコアの算出が上手く行われていた．また，会社名や商品名もアフィリエイトや製品の使用レポートなどの記事で出る傾向があるため，同じように比較的精度よく抽出ができた．一方で，中央省庁名については多くの場合ニュースの引用記事で出現するため，そのニュースの内容ごとに記事の内容が大きく変わる傾向があった．その結果全体として文書集合間の類似度が低くなり，抽出精度を低下する結果となった．

今回は 50 記事以上の文書に含まれる候補語句のみを算出の対象としたが，実際には正解となる省略語は全て 100 記事以上のブログ記事に含まれていた．この値は使用するブログ記事コーパスの規模にも依存するところであるので，固有名詞の正式名称を含む記事数の 10 分の 1 以下の記事にしか出ない語句は切り捨てる，といったような比率を利用した基準の方が望ましいと考えられる．

5. ま と め

ブログ記事集合を対象とした分析を行う際に，固有名詞が様々に省略されて用いられるという問題に対して，ブログ文書集合を利用した固有名詞の省略語の自動抽出手法の提案を行った．約 3500 万記事のブログ文書集合を用いることにより，固有名詞と一般に使われるその省略語のセットに対する，提案手法の精度評価を行った．その結果，F 値が最高で 69.6%となり，全体として再現率 75%までの範囲で適合率 65%前後を達成していることを確認した．

今回正解として使用した省略語では,『横浜ベイスターズ』に対する『横浜』など，略語として用いられるが，もっと広義の固有名詞としても用いられる可能性がある多義語も区別せずに扱われていた．今後ブログ分析向けに固有名詞の正式名称表記と省略語表記を利用していくに当たっては，多義語となる省略語が想定した固有名詞の省略として用いられているか否かを判別する手法が必要である．出現文書のジャンル推定技術などを活用することによって，として用いられている場合か否かを判別できるよう手法を拡張していく必要がある．

文 献

[1] 鄒々堀正幹，青江純一，" カタカナ異表記の生成および統一手法," NL, vol.94, no.5, pp. 33-40, 1993.
[2] 増山毅司，中川裕志，"Web データを利用したカタカナ異表記の自動獲得," 言語処理学会第 11 回年次大会予稿集, 2005.
[3] 長家利和，" 情報検索装置および方法," 特願平 9-274323, 1997.
[4] 近藤光正，乾健太郎，松本裕治，"Web 文書を利用した半教師あり用語抽出," 言語処理学会第 13 回年次大会予稿集, 2007.