ブログ文書集合を用いた省略語抽出手法の検討

Clipped Word Extraction using Blog Documents

関口 裕一郎♠ 佐藤 吉秀♦
川島 晴美♦ 奥田 英範♦

Yuichiro SEKIGUCHI Yoshihide SATO
Harumi KAWASHIMA Hidenori OKUDA

ブログの急速な普及により，人々の体験や感想といった生の声が数多くネットワーク上で発信されている．そのような中でブログから話題となっている事柄を抽出する，マーケティング情報抽出技術へのニーズが高まってきている．しかしブログ記事は口語的な表現で記述されるため，分析時に重要となる商品名等の固有表現が省略して表記され，分析精度を低下させる原因の一つとなっている．本論文では，固有表現の正式表記から，そ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的とし，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すむ．また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．

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.

1. はじめに

近年のブログの浸透により，多数の一般の人々の感想や体験を綴った記事がネットワーク上に発信されるようになってきた．またそれと時を同じくして，インターネットを介したショッピングが，一般層へ普及してきている．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口コミ情報として幅広いユーザに利用されており，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が，日々更新されるブログ記事集合において自社の関連する商品や分野がどのように語られているか強く関心を持つようになってきており，それを自動的に抽出するブログ分析技術へのニーズが高まってきている．

一方，商品に関連する話題を抽出するには，分析対象となる商品名やその競合商品の名称，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[1]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抜粋することで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．

表 1 ブログにおける異表記語のパターンごとの分布
Table.1 Number of clipped words appear in blog documents

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

2. 固有名詞の表記ゆれ

2.1 ブログ記事中の固有名詞の異表記パターン

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 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．

- **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．
- **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．

♠ 正会員 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 sekiguchi.yuichiro@lab.ntt.co.jp
♦ 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 {sato.yoshihide, kawashima.harumi, okuda.hidenori}@lab.ntt.co.jp

・**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．
・**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．また今回は『PlayStation』と『プレステ』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．その為各項目の合計は 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， 275 組で全体の 57.3%を占めた．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 組存在した．

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は，正式名称から省略語が作成されるルールを用いる手法と，大量の文書コーパス中での語句の出現状況から確からしい省略語を抽出する手法の大きく 2 つが存在する．

ルールを用いた手法としては，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．[2] 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し，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．また，正式名称からの省略語生成ルールを大量の文書コーパスを用いて学習し，それを用いて省略語の判定を行う技術も提案されている．[3]

コーパスを用いた手法としては，正式名称から簡単なルールを用いて省略語候補を複数生成したうえで，各省略語候補を含む文書が含む語句集合の類似度から確からしい省略語を抽出する手法が提案されている．[4]

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や[5]，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．[6]

3. 提案手法

省略語は元の固有名詞と同義語となるため，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，

・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
・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各候補語句の省略語らしさのスコアとして数値化する

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．上記提案手法はコーパスを用いた手法の一種であり，従来の手法に見られたルールによる省略語候補の絞り込みを行わず，文書中での語句の使われ方の傾向のみから省略語を特定する手法となる．

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 $S(w_{ne}, w_{cl})$ を式(1)のように定義する．

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

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となる．本手法では，

・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類似度
・各文書の構成語句が持つ意味属性を元に生成する概念ベクトル[7]の類似度をコサイン類似度で数値化した類似度

の２つのを類似度算出手法を用いる．

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

表２ 評価に使用した固有名詞の一部
Table.2 Example of proper nouns

固有名詞	省略語数
中日ドラゴンズ	2
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
松下電器産業	2
厚生労働省	1

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 $S(w_{ne}, w_{c1})$ 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

以上の省略語取得手法を，ウェブからのクローリングにより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． 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．省略語候補を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条件は，検索結果が 50 記事以下と設定した．

4. 評価実験

提案手法の有効さを，スポーツチーム名 12 語，企業名 4 語，中央省庁名 4 語の計 20 語の固有名詞について，その正解となる省略語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の例を表２に示す．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』)であり， 1 つの固有名詞あたりの正解省略語数は平均で 2.7 語となった．

算出する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．またこの閾値を変化させることにより得られる，適合率と再現率の関係を表すグラフを図 1 示す．

文書間の類似度に語句ベクトルを用いた場合，概念ベクトルを用いた場合のどちらにおいても，再現率 60 %前後の時に適合率 70 %前後となる結果になった．語句ベクトルを用いた場合，適合率 62.7 %，再現率 77.8 %のときに F 値が 69.4 で最大となる．また概念ベクトルを用いた場合，適合率 70.2 %，再現率 61.1 %の時に F 値が 65.3 で最大となる．再現率が低い部分では概念ベクトルの方が高い適合率となり，再現率が高い部分においては逆の傾向がみられた．

また表３～４に，『中日ドラゴンズ』『経済産業省』について語句ベクトルと概念ベクトルを用いて省略語スコアを算出した場合の上位６つの語句を，それぞれのスコアとともにを示す．

『中日ドラゴンズ』については，それぞれ正解である『中日』と『ドラゴンズ』，といった語句が上位に来ている．また，上位４語については語句ベクトルの場合も概念ベクトルの場合も，一部順位の入れ替わりはあるが同じ語句の組み合わせになっており，似た抽出傾向となっていることが分かる．

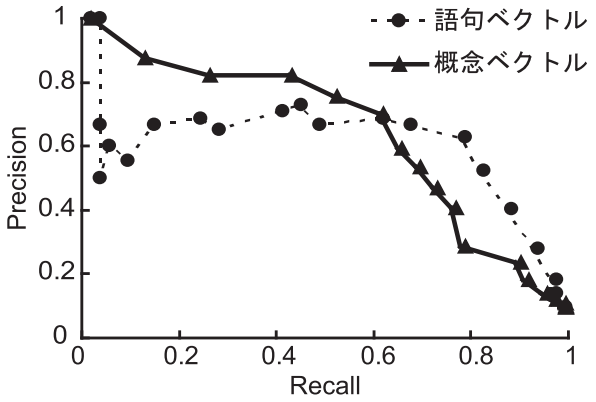

図１ 適合率と再現率の関係
Fig.1 Precision-Recall curve

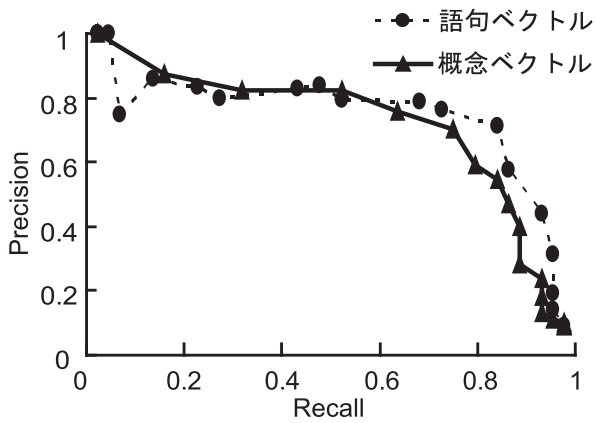

図２ スポーツチームのみでの適合率と再現率の関係
Fig.2 Precision-Recall curve in sports team names

一方『経済産業省』ついては，正解の『経産省』は語句ベクトルを用いた場合は 6 番目となり，概念ベクトルを用いた場合は 3 番目となっている．またどちらの場合も『中日ドラゴンズ』の場合に比べ，低いスコアが算出されていることが分かる．

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．

これは各固有名詞のブログ記事中での扱われ方によるものと考えられる．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，そのような記事中では，例えば「ホームラン」や「凡退」といった，共通して現れる語句が存在し，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も専門用語が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．

表 3 『中日ドラゴンズ』の略語スコア上位 6 語
Table.3 Top6 clipped words of “Chunichi Dragons”

略語	スコア(語句ベクトル)	略語	スコア(語句ベクトル)
ドラゴンズ	0.241	中日	0.794
中日	0.165	ドラゴンズ	0.761
日ドラゴンズ	0.110	日ドラゴンズ	0.665
ドラズ	0.059	ドラゴ	0.591
ゴンズ	0.058	中ドラ	0.560
ゴズ	0.055	ドラゴン	0.513

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．

また語句ベクトルを用いた場合と概念ベクトルを用いた場合において，再現率と適合率の関係に異なる傾向が見られた．これはブログでは記事ごとに文体が大きく異なるためと考えられる．概念ベクトルを用いた場合は，語句の意味属性を見て類似度を算出するので，似た内容の文書で異なる言葉使いをしている場合でも適切に類似度が算出される．この為再現率が低い範囲で高い適合率を示すと考えられる．一方で使用される語句のジャンルが近い場合でも類似しているとみなす場合がある為，再現率が高い範囲では誤抽出が増え，適合率を下げる結果をもたらしたと考えられる．

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. まとめと今後の課題

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 69.4 となり，全体として再現率 60％前後の範囲で適合率 70％前後を達成していることを確認した．

今回の手法においては，略語か複数存在するスポーツチームでは比較的精度か高く抽出できたが，中央省庁名など略語がほぼ一つに絞られる分野においては精度が出なかった．略語が一つに絞られるような語句は，ルールベースの省略語抽出によって精度の良い抽出が可能な対象である．従って，対象によってアルゴリズムを切り替えるような工夫により，更なる精度向上を図っていきたい．

また今回使用した正解には，『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より広義の固有名詞としても用いられる語句も含まれている．そのため作成した略語辞書を用いた固有名詞抽出を行う際には，各文書の文脈を分析して，これらの語句が省略語として使われているのか他の意味の固有名詞として使われているのかを，判別するような技術の構築が必要と考えられる．

表 4 『経済産業省』の略語スコア上位 6 語
Table.4 Top 6 clipped words for “Ministory of Economy, Trade and Industory”

略語	スコア(語句ベクトル)	略語	スコア(語句ベクトル)
産業	0.052	経済産業	0.424
経済省	0.049	産業	0.417
経済	0.045	経産省	0.406
経産	0.044	経産	0.403
産業省	0.043	産業省	0.396
経産省	0.043	経済省	0.387

[文献]

[1] 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.

[2] 長家利和，” 情報検索装置および方法,” 特願平 9-274323，1997.

[3] 村山紀文，奥村学,“Noisy-channel modelを用いた略語自動推定,”言語処理学会 第12回年次大会，pp.763-766，2006.

[4]酒井浩之，増山繁,“コーパスからの名詞と略語の対応関係の自動獲得,”言語処理学会 第9回年次大会，pp.226-229，2003.

[5] 獅々堀正幹，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.

[6] 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

[7] 別所 克人，古瀬 蔵，片岡 良治，”単語と意味属性との共起に基づく概念ベクトル生成手法，”人工知能学会第20会全国大会論文集，2006.

関口 裕一郎 Yuichiro SEKIGUCHI
NTT サイバーソリューション研究所所属．2004 年東京大学大学院情報理工学系研究科修士課程修了．同年日本電信電話（株）入社．現在インターネットにおける情報抽出の研究開発に従事．情報処理学会，日本データベース学会各会員．

佐藤 吉秀 Yoshihide SATO
NTT サイバーソリューション研究所所属．2000 年京都大学工学部電気電子工学科卒業．2002 年同大学院情報学研究科システム科学専攻修了．同年日本電信電話（株）入社．自然言語処理の研究に従事．情報処理学会会員

川島 晴美 Harumi KAWASHIMA
NTT サイバーソリューション研究所主任研究員．1990 年山梨大学大学院工学研究科修士課程修了．同年日本電信電話（株）入社．現在インターネットからの話題情報抽出技術の研究開発に従事．電子情報通信学会会員．

奥田 英範 Hidenori OKUDA
NTT サイバーソリューション研究所主幹研究員．1988 年東京大学大学院工学系研究科修士課程修了．同年日本電信電話(株)入社．1994 年スタンフォード大学コンピュータ科学科修士課程修了．現在 CGM における情報抽出の研究開発に従事．電子情報通信学会，映像情報メディア学会各会員．