Wikipediaの表記特徴を利用した別称コーパス生成ツールの開発

山西良典† 福本淳一†
立命館大学情報理工学部メディア情報学科
{ryama,fukumoto}@media.ritsumei.ac.jp

概要 本稿では，Wikipedia の構造特徴および表記特徴を利用した別称コーパス生成ツールを開発した．正式名称の他に別称を持つ知識は多く，特に Web 上では別称での記述が多い．提案ツールは，正式名称と「略称」「愛称」「通称」それぞれが対応づいたコーパスを，Wikipedia の構造と表記特徴を利用して生成する．ダイナミックに編集・更新され，固有名詞に関する記事が多く存在する Wikipedia を情報源とすることで，日々増加する固有名詞についても対応したコーパスの自動生成が可能となる．評価実験の結果，生成された各コーパスは自動抽出されたコーパスとしては非常に高い精度で別称が収集されていることを確認した．正式名称と別称が対応付いたコーパスを生成することで，Web 上で別称を用いて記述された意見・評判の取得が可能となるだけでなく，正式名称からの別称推定研究における学習データとしての応用も期待される．

キーワード 別称，コーパス，Wikipedia，リンク構造，オープンデータ

1 はじめに

ウェブ上の知識を利用する研究の多くでは，情報を得たい対象の名称を検索エンジンなどに入力することで，関連情報や評判情報を取得している [1, 2, 3, 4]．このとき，入力される名称は正式名称であることが一般的であるが，正式名称とは別に「略称」「愛称」「通称」といった「別称」で呼ばれることが多い対象も存在する．例えば,「日本放送協会」は「NHK」,「指原莉乃」は「さっしー」,「横浜国際総合競技場」は「日産スタジアム」のような別称での記述が多い．これは，略称が正式名称を略すことで字数が削減されていることや，対象についての印象などの感性的な記述をする場合には親しみを込めた愛称を用いることが多いためであると考えられる．

別称は，正式名称が言い換えられた形であり，大別すると「略称」「愛称」「通称」の 3 つが存在する．略称は，ある語句の一部分を抽出して省略した語句であり，字数制限が設けられた SNS 上で頻繁に用いられる．愛称は，ある語句に対する親しみを込めた呼称であり，印象や感想などを記述する際に用いられることが多い．通称は略称や愛称を包括し，一般的に知られ対象を呼称する場合に広く用いられる別称である．ウェブ上からの知識を利用しようとした場合，例えば感性的な意見を取得するためには正式名称と愛称との対応づけが必要となる一方で，略称の自動推定手法のための学習データとしては正式名称と略称の対応が必要となる．そこで，本稿の提案ツールでは，これらの 3 種の別称を選択的に抽出し，正式名称と別称が対応づいたコーパスを生成する．

一般名詞の別称であれば，あらかじめ正式名称と別称を対応付けたコーパスを用意しておくことで対応できる．しかしながら，日々増加していく固有名詞については別称を静的なコーパスとして用意することは難しく，動的に更新される固有名詞の別称コーパスが必要となる．既存研究では，発音を基にして正式名称からの略語推定 [5] や Web 検索エンジンを利用した略語推定 [6] が報告されており，これらの手法を用いることで広義には動的な別称が獲得出来ると考える．しかしながら，これらの研究では略称推定器を構築するために一般名称と略語が対応づいた学習データが必要となり，現在は全て人手で作成した辞書が用いられている．また，これらの手法によって推定された別称が実社会において一般的に用いられる別称であるとは限らない．

本稿では，上記の背景のもとで，Wikipedia の構造特徴および表記特徴を利用した別称コーパスの生成ツールの開発を行った．提案ツールは，一般ユーザが動的に更新する Wikipedia 上の知識を利用することで，動的に更新され，一般的に用いられる可能性が高い別称のコーパス生成を実現する．また，既存研究で頻繁に扱われている略語のみならず，愛称や通称，別名といった別称を選択的に抽出可能である．

2 別称の使われた方

近年，ソーシャルネットワークサービス（SNS）上では様々な対象についての意見が交わされている．SNS では，主観的，直感的な意見が多く，対象についての感性的な評価を取得するうえで有用なデータと考えられる [7]．しかしながら，文字数に制限がある SNS 上での感性的な意見の多くは，正式名称ではなく略称や愛称によって記述されていることが多く，正式名称は広告的な投稿にのみ使われていることが多い．

例えば，正式名称「ももいろクローバー Z」の別称（愛

表 1 「ももいろクローバー Z」と「ももクロ」の Twitter 上での使われた方の違い．100 件中の件数

	ももいろクローバー Z	ももクロ
感性的な意見の投稿	2	70
広告的な投稿	98	30

- @kkkkmek : ももクロキタ――――――(∀)――――――!!!!! http://t.co/WFesAHI0Vr Wed Oct 30 09:48:07 +0000 2013
- @gintamahm : ももクロ熱が上がって来た Wed Oct 30 09:47:16 +0000 2013
- @mcz5_r : ももクロの新曲いいわーヽ(*´∀`)ノ Wed Oct 30 09:46:48 +0000 2013
- @maticaqp : (o^)o 朝から夜まで、ももいろクローバー Z http://t.co/UI0csQtooi Wed Oct 30 08:45:32 +0000 2013

図 1 感性的な意見の tweet 例

称）は「ももクロ」である．正式名称と別称について，同時刻にそれぞれ Twitter 上で検索し，tweet を 100 件ずつ取得した（取得日 2013 年 10 月 30 日）．取得した tweet について，人手で感性的な意見であるか，広告的な投稿であるかを分析した．分析結果を表 1 に示す．ここで，感性的な意見の投稿とは図 1 に示すように意見や感想，ユーザの状況が記述された tweet を指し，広告的な投稿とは図 2 に示すように商品説明や宣伝が記述された tweet を指す．

表 1 から，正式名称「ももいろクローバー Z」が使われている tweet のほぼ全てが広告的な投稿であるのに対して，別称の「ももクロ」が使われている tweet では感性的な意見が記述されている割合が極端に高いことが見て取れる．感性的な意見を投稿する際には親しみや愛着を込めるため愛称が使われ，広告的な意見では正式名称を用いることで正確に商品情報を伝えようとしていると考えられる．本結果より Web を情報源とする場合，正式名称だけでなく愛称や略称などの別称を用いることで，より感性的な意見が得られる可能性が高いことが示唆された．

3 Wikipedia の構造および表記特徴

Wikipedia は誰もが自由に編集可能なウェブ上の百科事典であり，87 万件以上の記事が存在し現在も記事数は増加している．Wikipedia には人名，地名，組織名など新語を含む様々な対象について，そのものの意味や関連情報などが記述されており，別称についての記述も多く存在する．

- @ZmomoZcloZ : ももいろクローバー Z 入口のない出口 (初回限定盤 A) [CD+Blu-ray, Limited Edition] http://t.co/vPBCk4gXgj #momoclo ももクロ Wed Oct 30 10:24:06 +0000 2013
- @follow_mex2x2 : ももいろクローバー Z 応援委員会 の ももいろクローバー Z × galaxxxy 行くぜっ！怪盗ヒョウジョパーカー【百田夏菜子】Kanakoo ↑ Red を Amazon でチェック！ http://t.co/hDlHjiI95w #autofollow #sougofollow Wed Oct 30 10:23:21 +0000 2013
- @PleasureNiiza : 【プレジャNiiza】「閉店セール 20 % OFF 中」ももいろクローバー Z 地域最大級の品揃え！まずはご来店下さい。#プレジャ #ももいろクローバー Z #ももクロ Wed Oct 30 10:10:23 +0000 2013
- @tomy9664 : ももいろクローバー Z 責任編集 『ももクロぴあ vol.2』[ムック] http://t.co/Y5K4q0Yl6a 発売予定日は 2013 年 7 月 2 日特典：ももいろクローバー Z のライブイベント「Summer Dive 2013」の読者限定スーパーリザーブシート予約応募ハガキ Wed Oct 30 09:55:58 +0000 2013

図 2 広告的な意見の tweet 例

3.1 リダイレクトページ

Wikipedia では，リダイレクトページと呼ばれるページが，いくつかの項目に用意されている．リダイレクトページは正式名称とは異なる入力に対して，正式名称を項目とするページへ転送するために用意されている．例えば，前述の「ももいろクローバー Z」について見てみると，愛称の「ももクロ」と入力することで正式名称である「ももいろクローバー Z」のページへと転送される．

リダイレクトページが用意されている項目は，ユーザが正式名称以外を入力する可能性が高い項目と考えられ，一般的に認知度の高い別称を有している項目と捉えられる．本稿の提案ツールでは，リダイレクトページを有する項目を対象とし，ユーザが正式名称以外で呼称することが多い項目についての別称を抽出する．

3.2 アブストラクトと基本情報の表

Wikipedia は，不特定多数のユーザが自由に編集しているが，その表記には特徴が見られる．Wikipedia 上に存在するほぼ全てのページにおいて，ページ上部に表題

堀北 真希（ほりきた まき、1988 年 10 月 6 日 - ）は、日本の女優、タレント。本名非公開。愛称は、**真希ちゃん**、**まきまき**、**ホマキ**など。東京都清瀬市出身。スウィートパワー所属。スリーサイズは B78、W58、H83cm。特技はピアノ、料理。

図 3　Wikipedia 上の堀北真希についてのアブストラクト文

項目を説明するアブストラクト文や基本情報の表が記述されている．

アブストラクトや基本情報の表には，表題項目の定義文や関連情報（例えば，人物であれば出身地や所属，建造物であれば所在地）などが記載されている．アブストラクトには「愛称/略称/通称は〜」といった記述や，基本情報の表には「愛称/略称/通称：〜」といった項目が用意されていることがある．これらの単語を手掛かりとすることで，表題項目を正式名称とする別称を取得する．

4　別称抽出ツール：HAP

本稿で提案する別称抽出ツール（Hypocorism and Abbreviated/Popular name extraction tool：HAP）では，Wikipedia のダンプデータを基に以下の手順で別称を抽出する．

1. リダイレクト元とリダイレクト先が対応づいたリストを作成
2. リダイレクト先ページから表記特徴を手掛かりとして別称候補を抽出
3. 抽出した別称候補とリダイレクト元ページの項目名を照合

まず，Wikipedia のダンプデータからリダイレクト元とリダイレクト先が組となったリストを生成する．リダイレクト関係を基にリストを生成する理由は，リダイレクト元は別称，リダイレクト先は別称となっている可能性が高いためである．このとき，リダイレクト先としてページ位置までを指定しているリダイレクト元は不採用とした．これは，ページ位置を指定したリダイレクト元の表題項目は，リダイレクト先の表題項目と一致しない例が複数見られたためである．

次に，リダイレクト先のページ中のアブストラクトおよび基本情報の表から表記特徴を基に別称を抽出する．アブストラクト中で,「略称は」「愛称は」「通称は」を含む文中で太字の記述，あるいは，カギ括弧またはシングルクォーテーション 2 つ以上によって囲まれた文字列を別称候補として抽出した．例えば，図 3 に示した堀北真希のアブストラクト文からは愛称として「真希ちゃん」「まきまき」「ホマキ」が抽出される．このとき，別称に付与された脚注やリンクは除去し，参照文字については本来の文字に復元した形で出力を行う．また，別称の後ろの丸括弧内によみがなが記述されている事例も確認されたため，提案ツールでは別称の後ろの丸括弧および丸括弧で囲われた文字は除去した．

表 5 別称コーパスのそれぞれの適合率評価．数値は%．

	略称	愛称	通称
適合率	0.857	0.924	0.980

そして，抽出した別称候補をリダイレクト元の項目名と照合することで，より一般的に用いられる可能性が高い別称の抽出をねらう．これは，リダイレクト元ページはユーザが入力する可能性が高い別称を項目名としているためである．

4.1　別称コーパス例

2013 年 9 月 6 日時点での Wikipedia の最新ダンプデータに対して，提案ツールを用いてコーパスの生成実験を行った．その結果，略称，愛称，通称についてそれぞれ 1422 件, 459 件, 343 件が抽出された．提案ツールによって抽出された略称，愛称，通称の一部をそれぞれ表 2, 3，4 に示す．

生成された略称コーパス，愛称コーパス，通称コーパスについて，それぞれ妥当性評価実験を行った．評価実験では，母比率 0.1，標準誤差 0.05 で信頼度 95%を満たすサンプル数を算出し，略称，愛称，通称についてそれぞれ 126，106，99 サンプルを取り出して評価した．

本稿では，20 代の評価者を 2 名用意し，2 名の評価者が共通して不適切とした項目を誤抽出として，適合率を算出した．表 5 に，評価実験の結果を示す．同表から，全てのコーパスについて，85%以上の適合率で別称が抽出されていることが見て取れる．人手を加えずに，自動的に抽出されたコーパスの性能としては高い適合率が示された．これは，提案ツールが表記特徴から抽出された別称候補を，リダイレクト元の項目名と照合した上で出力したためと考えられる．リダイレクト元の項目名との照合は，別称候補として抽出された文字列が wikipedia ユーザがページ検索時に利用する可能性が高い別称であるかを検証することに相当する．

4.2　考察

本節では，評価実験結果を基に提案ツールによって作成されたコーパスについて詳細に考察する．まず，適切に抽出された例について，それぞれのコーパス毎に考察する．

略称コーパスでの抽出例

略称コーパスでは,「ワードプロセッサ − > ワー

表 2 抽出された略称の例

正式名称	略称	正式名称	略称	正式名称	略称
はるやま商事	はるやま	ELLEGARDEN	エルレ	韓国取引所	KRX
03 式中距離地対空誘導弾	中 SAM	富士通モバイルコミュニケーションズ	富士通モバイル	ワードプロセッサ	ワープロ
個人情報の保護に関する法律	個人情報保護法	ロンドンオリンピック・パラリンピック組織委員会	LOCOG	早稲田大学高等学院・中学部	早大学院, 早高院
米国国家規格協会	ANSI	大日本除虫菊	KINCHO, 金鳥	スペイン社会労働党	PSOE
超高温材料研究所	JUTEM	UFJ ホールディングス	UFJHD	メタルスラッグ	メタスラ
関西独立リーグ	KANDOK	さくらシュトラッセ	さくラッセ	静岡第一テレビ	だいいちテレビ
天童市立第四中学校	天四中	スターバックス	Starbucks	テレビ埼玉	テレ玉
ニコニコ生放送	ニコ生	魔法少女まどか☆マギカ	まどマギ	スーパーマリオブラザーズ	スーマリ
三菱 UFJ フィナンシャル・グループ	MUFG	ときめきメモリアル	ときメモ	ムヒョとロージーの魔法律相談事務所	ムヒョロジ, ムヒョ

表 3 抽出された愛称の例

正式名称	愛称	正式名称	愛称	正式名称	愛称
青田典子	バブル青田	少年陰陽師の WEB ラジオ	孫ラジ	第 66 回国民体育大会	おいでませ!山口国体
林家木久扇	木久ちゃん	ひらかたパーク	ひらパー	名古屋ガイドウェイバスガイドウェイバス志段味線	ゆとりーとライン
浅草花やしき	花やしき	瞳と光央の爆発ラジオ	爆ラジ	高浜市やきものの里かわら美術館	かわら美術館
安倍麻美	あさみん	長岡移動電話システム	FM ながおか	喜多村英梨	キタエリ
ロサンゼルス	L.A.	松山ケンイチ	松ケン	サッカー日本女子代表	なでしこジャパン
横浜国際総合競技場	日産スタジアム	宮澤佐江	さえたむ	山本梓	あずあず
酒井法子	のりピー	浅草花やしき	花やしき	愛知高速交通東部丘陵線	リニモ, Linimo
さよなら絶望放送	SZBH	竹達・沼倉の初めてでもいいですか?	初ラジ	大篠津町駅	砂かけばばあ駅
大阪ターミナルビル	サウスゲートビルディング	東京臨海副都心	レインボータウン	滝沢乃南	のなみん

プロ」や「メタルスラッグ − > メタスラ」「ELLEGARDEN − > エルレ」のように正式名称または読み方の文頭のみを用いた略称，正式名称の一部を抽出した「宇都宮地方裁判所 − > 宇都宮地裁」「岩手県立盛岡第一高等学校 − > 盛岡一高」「ときめきメモリアル − > ときメモ」のような略称が抽出された．また,「シティックスカード − > CITIX」「アメリカ大気研究センター − > NCAR」「米国国家規格協会 − > ANSI」「ザ・キング・オブ・ファイターズ − > KOF」のように英語表記の単語のイニシャルを用いた略称も多く見られた．その他には,「この中に 1 人、妹がいる! − > 中妹」や「僕は友達が少ない − > はがない」「もし高校野球の女子マネージャーがドラッカーの『マネジメント』を読んだら − > もしドラ」といった特殊な略称についても抽出された．

愛称コーパスでの抽出例

愛称としては,「山本梓 − > あずあず」や「酒井法子 − > のりピー」「河西智美 − > とも〜み」のように人名の一部を抽出したものを定型的に変化させたものが多く見られた．また,「にゃんにゃん丸 − > にゃん丸」や「ひらかたパーク − > ひらパー」のような略称が愛称として用いられているものもあった．これらは，正式名称の一部を利用して生成されている愛称パターンといえる．一方で,「サッカー日本女子代表 − > なでしこジャパン」や,「牛久市コミュニティバス − > かっぱ号」「大篠津町駅 − > 砂かけばばあ駅」のように正式名称と愛称の間で共有される文字列が存在しない愛称のパターンも存在した．これらの愛称については，機械学習による別称推定では推定不可能な事例と考える．

表 4 抽出された通称の例

正式名称	愛称	正式名称	愛称	正式名称	愛称
ミュージックステーション	M ステ	筑波大学附属駒場中学校・高等学校	筑駒	法然	黒谷上人, 吉水上人
紳助社長のプロデュース大作戦!	プロデュース大作戦!	リヨン国立高等音楽・舞踊学校	CNSMDL	西日本電信電話	NTT 西日本
大阪大学生活協同組合	阪大生協	ニンテンドードリーム	ニンドリ	ポニーキャニオン	ポニキャン
全国高等学校野球選手権大会	夏の甲子園, 夏の高校野球	J リーグカップ	ナビスコカップ	機動捜査隊	機捜
たばこの規制に関する世界保健機関枠組条約	たばこ規制枠組条約, たばこ規制枠組み条約	イオンタウン千種	イオン千種	天才てれびくん	天てれ, 天テレ
飛田遊廓	飛田新地	開運!なんでも鑑定団	鑑定団, なんでも鑑定団	七対子	チートイ
滝川クリステル	滝クリ	SmaSTATION!!	スマステ	奇跡体験!アンビリバボー	アンビリーバボー
名古屋テレビ放送	名古屋テレビ, メ〜テレ	日本興業銀行	IBJ	徳島バス	徳バス

通称コーパスでの抽出例

通称は，略称や愛称などを包括しているため，通称のパターンには様々なものが存在した．例えば,「平頼盛 − > 池殿」や「法然 − > 黒谷上人」「熱傷 − > 火傷」のように，対象が呼称される渾名や正式名称に対してより一般に知られている名称などが通称として抽出された．また,「ミュージックステーション − > M ステ」のようにカタカナ表記を本来の英語に復元した上でイニシャルを用いたものや,「SmaSTATION!! − > スマステ」のように英語表記の発音をカタカナ表記して一部を抽出したものなども見られた．一方で,「大阪大学生活協同組合 − > 阪大生協」や「機動捜査隊 − > 機捜」のように略称に多く見られるパターンでの表記された通称も多く存在した．

次に，誤抽出について，それぞれのコーパス毎に考察する．

略称コーパスでの誤抽出

略称コーパスでの誤抽出（126 サンプル中 18 項目）では,「意味を考慮した略」「カタカナ英語変換」がそれぞれ 7，4 項目と多数を占めた．意味を考慮した略の例としては,「行政手続等における情報通信の技術の利用に関する法律 − > 行政手続オンライン化法」や,「為公会 − > 麻生派」「濃飛乗合自動車 − > 濃飛バス」などが挙げられる．これらの略称は人間が対象についての意味を基に略称を連想しているパターンであり，正式名称そのものから文字列を選択・短縮した一般的な略称とは異なる．また,「カタカナ英語変換」の例としては「スカイ・エー − > スカイ・A」や「日立オムロンターミナルソリューションズ − > Leadus」などが挙げられる．これらは，カタカナの読みを英語表記に変えることで文字数を削減したパターンである．その他のパターンとしては,「東京録音現像 − > 目黒現像所」のように対象が存在する場所を用いて略称としているものや,「サルゲッチュ − > ピポサル」のように対象中の登場人物を用いて表現しているもの,「牛乳石鹸共進社 − > 牛乳石鹸」のように略されているものの略称が曖昧であるものなどが誤抽出として評価された．

愛称コーパスでの誤抽出

愛称コーパスの評価実験では，106 サンプル中 8 項目が誤抽出として判断された．愛称コーパスでの誤抽出には,「まつながひろこ − > 松永裕子」や「松本秀夫 − > 松本ひでお」といった「かな漢字変換」を愛称としているもの,「京都放送 − > KBS 京都」や「福島放送 − > KFB 福島放送」のように正式な別名を愛称として抽出しているものがあった．その他には,「ソフィア・コワレフスカヤ − > コヴァレフスカヤ」や「国立アメリカ・インディアン博物館 − > NMAI」のように略称が愛称として用いられているパターンも存在した．また，スポンサーが名称権を獲得して愛称としている「広島広域公園陸上競技場 − > エディオンスタジアム広島」や，対象が存在する場所と使用用途を愛称としている「兵庫県立淡路佐野運動公園 − > 淡路球場」といったパターンも存在した．

通称コーパスでの誤抽出

通称については 0.980 と非常に高い適合率（99 サンプル中 2 項目が誤抽出）となったが，これは通

称が略称や愛称を含む包括的な別称であるためであると考えられる．評価実験において通称コーパス中で誤抽出と判断された項目は「リヨン国立高等音楽・舞踊学校 − > CNSMDL」と「滝川クリステル − > 滝クリ」であり，どちらも略称であった．これらについて，評価者に不適切と判断した理由を聴取したところ，これらは通称として一般的に普及した呼称ではないと判断したためであるとの回答を得た．

別称のタイプ別での評価では不適切と判断された誤抽出パターンの中には，別称としては十分有用性が高いと考えられるものも存在した．例えば,「行政手続等における情報通信の技術の利用に関する法律 − > 行政手続オンライン化法」や「サルゲッチュ − > ピポサル」などは，一般的に認識されている略称の形式とは異なっているため，評価実験では不適切な略称と判断された．しかしながら，Web 上で「行政手続オンライン化法」,「ピポザル」について検索を行うと，それぞれ「行政手続等における情報通信の技術の利用に関する法律」や「サルゲッチュ」についての情報を取得することが出来る．これらの意味や内容から生成された略称については，表記や発音を用いた略称推定手法では獲得することが出来ない略称であり，自由に編集された Wikipedia を情報源として用いるからこそ抽出された略称と考える．

5 おわりに

本稿では，Wikipedia の構造と表記特徴を利用した別称の自動生成ツールを開発した．提案ツールでは，Wikipedia 上でリダイレクトページが用意されている項目（リダイレクト先）は別称で検索される可能性が高い項目であると考え，リダイレクト先ページ内のアブストラクト文中で略称について記述されている文からパターン照合により別称候補となる文字列を抽出した．そして，抽出した別称候補をリダイレクト元ページと照合することで，正確な正式名称と別称の組み合わせの抽出をねらった．

提案ツールを用いて別称コーパスを生成し評価したところ，提案ツールによって抽出された別称コーパスは機械的に生成されたコーパスとしては非常に高い適合率が確認された．提案ツールによって抽出された別称には，機械学習では推定が難しいと考えられる別称も多く確認できた．

Web 上で任意の対象について情報を取得する場合に，提案ツールによって生成された別称コーパスを参照することで，別称によっても情報を取得することが可能となる．そのため，2 節で示したような対象についてのより感性的な意見の収集が期待される．また，別称推定手法開発における学習データとしての応用も考えられ，学習用コーパスを人手で作成する手間の削減に繋がると考える．今後は，コーパス中によみがなや異表記，対象の定義情報などの情報も追加し，コーパスとしての有用性を高めていく．

謝辞

本稿の執筆にあたり，角野翔太氏の協力を得た．記して謝意を表す．また，本研究は一部，中部電気利用基礎研究振興財団の助成のもと行われた．

参考文献

[1] 桝井文人，福本淳一，荒木健治，“比喩解釈を目的とする world wide web を利用した属性値の適合性判定手法とそのフィードバック,” 電子情報通信学会論文誌，vol.J89-D，no.4，pp.860–870，2006.

[2] R. Yamanishi, J. Fukumoto, and F. Masui, “Semantical-coordinate terms detection from hierarchical knowledge using web snippets,” Procedia Computer Science, vol.22, pp.1276–1284, 2013.

[3] 松尾豊，友部博教，橋田浩一，中島秀之，石塚満，“Web 上の情報からの人間ネットワークの抽出,” 人工知能学会論文誌，vol.20，no.1，pp.46–56，2005.

[4] 大島裕明，小山聡，田中克己，“Web 検索エンジンのインデックスを用いた同位語とそのコンテキストの発見,” 情報処理学会論文誌，vol.47，no.19，pp.98–112，2006.

[5] 和田健太，近山隆，横山大作，三輪誠，“素性にモーラとシラブルを用いた略語の自動推定,” 情報処理学会研究報告. 自然言語処理研究会報告，vol.2009，no.36，pp.67-72，2009.

[6] 村山紀文，奥村学，“Web 情報を利用した確率モデルによる略語推定,” 情報処理学会研究報告. 情報学基礎研究会報告，vol.2008，no.4，pp.93-100，2008.

[7] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” Journal of Computational Science, vol.2, no.1, pp.1-8, 2011.