Wikipedia を用いた人物別名の抽出と人物判別のためのラベル付与

齊藤 大樹[1]　　分部 裕二[2]　　内田 理[2]

[1]東海大学工学研究科情報理工学専攻

2bdrm013@mail.tokai-u.jp

[2]東海大学情報理工学部情報科学科

o-uchida@tokai.ac.jp

1. はじめに

芸能人やスポーツ選手などの著名人の多くには，あだ名や愛称等の別名が存在する．テレビや新聞，Web などのメディアで著名人が報道される際は，正式名称のみならず，別名が使用されることも多い（例えば，元 AKB48 の前田敦子さんは「あっちゃん」という愛称が頻繁に利用される）．そのため，ある著名人に関する情報をインターネット上で漏れなく検索，収集するためには，このような人物名の記述のゆらぎを考慮する必要が生じる．加えて，一つの別名が複数の著名人に該当する場合，インターネット上の検索結果から目的の人物の情報のみを発見することは困難である．そのため，別名から特定の著名人を識別することも重要である．以上のことから，正式名称と別名が対応付けられたデータベースの作成が必要であると考えられる．しかし，ファンやメディアによって時々刻々と新しく別名が生み出されるため，定期的にデータベースの更新が必要である．また，ある人物に対して複数の別名が与えられるケースも多く，その全てを収集することは容易ではない．そのため，手動によるデータベースの作成や，その維持管理は現実的ではない．

ところで，ある対象についての知識，情報を取得する手段の一つとして，百科事典の利用が挙げられる．オンライン百科辞典としては Wikipedia が最も大規模であり，日本語版だけでも約 89 万件の記事が存在する（2014 年 1 月 22 日現在）．著名人に関する情報も多く記載されており，例えば，日本人の元メジャーリーグベースボール選手である松井秀喜氏についての記事（http://ja.wikipedia.org/wiki/松井秀喜）では，彼の出身地，経歴などのほか，彼の愛称（ゴジラ）も記述されている．また，Wikipedia は多くのユーザによって編集が行われるという特徴を有しており，定期的に更新がなされている．

そこで本研究では，Wikipedia を用いて人物の別名を自動的に抽出し，また人物判別のためのラベル付与を行う手法を提案する．

2. 関連研究

Hokama らは，ある人物の別名を記述する際に “別名「こと」正式名称” の表現が多く用いられることに着目した別名抽出手法を提案している [1]．Hokama らの手法では，まず「こと」と別名を取得したい著名人の正式名称で構成された “こと正式名称”（例えば，“こと松井秀喜” ）を用いて Web 検索を行う．その後，形態素解析により “こと正式名称” の直前に出現する文字列を取得し，別名の候補としている．また Bollegala らは，Hokama らが利用した “こと正式名称” だけでなく，より多くの表現方法を利用した別名抽出手法を提案している [3]．Bollegala らの方法では，まず別名が明らかとなっている正式名称とその別名のデータセットを用いて，“正式名称 * 別名” ，及び “別名 * 正式名称” で Web 検索を行う．その検索結果から「*」に該当する文字列を抽出し，パターン候補とする．それらパターン候補がどの程度，別名候補を取得可能かを評価し，パターンを決定する．最後に，生成したパターンを利用して別名候補を取得している．

しかしこれらの先行研究の問題点の一つに，別名とそれ以外との切れ目の判定精度が低いことが挙げられる．日本語は英語等の言語と異なり，単語と単語の間に明確な区切り（スペース）が存在しない．そのため，一つの単語を正確に抽出することが困難であり，抽出した別名にノイズが含まれる可能性がある．この問題の解決方法としては形態素解析の利用が考えられるが，別名は形態素解析器の辞書に未登録（未知語）であることが多く，正しく抽出できない可能性が高い．そこで本研究では，Wikipedia における特徴的な別名記述パターンを用いることにより，この問題の解決を試みる．

3. 提案手法

3.1 Wikipedia ページの取得

Wikipedia 財団では，誰でも再配布や再利用が出来るよう，コンテンツのデータを XML 形式で提供している．本

研究では，このデータから必要な情報を取得する．なお，XML 特有のタグや，Wikipedia ページのリンクを表す 2 つの連続する角括弧はノイズとなるため，あらかじめ削除しておく．

3.2 特徴キーワードを含む文の取得

「呼ばれ」や「称され」など，特定の単語（以後，特徴キーワードと呼ぶ）を含むパターンの前後に別名が記載されていることが多い．そこで，特徴キーワードを利用して別名が記載されている可能性が高い文を取得する．なお，予備実験の結果より，特徴キーワードは表 1 のように設定した．

3.3 別名記述パターンを用いた別名候補の抽出

Wikipedia 内で別名が記載される際，以下のような 2 つの特徴が見受けられる．

- 別名が記載されている文中にある特徴キーワードには，それぞれ対応した接頭辞，もしくは接尾辞が存在する．多くの別名はそれら接頭辞の前，もしくは接尾辞の後に記載されている．
- クォーテーション，またはかぎ括弧で括られた文字列は別名である可能性が高い（例えば Wiki 文法では，ダブルクォーテーション（''）で文字列を括ることは文字列を斜体にすること，トリプルクォーテーション（'''）で文字列を括ることは，文字列を太字にすることを意味する）．

提案手法では，これらの特徴を別名記述パターンとして設定し，別名候補の抽出を行う．それぞれの特徴キーワードの接頭辞，及び接尾辞は，予備実験の結果より，表 1 のように設定した．抽出手法を以下に示す．

1. 取得したすべての文に対し，以下のルールを適用する．
   (a) 表 1 に記す特徴キーワードと接頭辞の組み合わせが存在する時，接頭辞の前にある，クォーテーション，かぎ括弧，またはその両方で括られた文字列を抽出する（図 1 (a)）．該当する文字列がなければ，抽出は行わない．
   (b) 表 1 に記す特徴キーワードと“ = ”以外の接尾辞の組み合わせが存在する時，接尾辞の後にある，クォーテーション，かぎ括弧，またはその両方で括られた文字列を抽出する（図 1 (b)）．該当する文字列がなければ，抽出は行わない．
   (c) 表 1 に記す特徴キーワードと接尾辞“＝”の組み合わせが存在する時，その後に存在する文字列を抽出する（図 1 (c)）．文字列の間に句読点が存在する場合は，その句読点を境に分割し，分割後のそれぞれの文字列を抽出結果とする．
2. 抽出した文字列を別名候補と決定する．ただし，抽出した文字列に重複するものが存在する場合は，1 つを除き削除する．

…優勝記者会見で石川本人も**「ハニカミ王子」のニックネーム**についてコメントし…

(a) 接頭辞の場合

…に[[国民栄誉賞]]を受賞。**愛称は「'''ゴジラ'''」**==経歴 ===== プロ入り前 ==…

(b) 接尾辞の場合

| Alias = ひーちゃん、ひとみん、ひっと、ひーすけ、ガチャピン先輩

(c) “＝”の場合

図 1　パターン抽出の例

表 1　別名記述パターン

接頭辞	特徴キーワード	接尾辞
と	呼ばれ	
とも		
と	称され	
とも		
の	異名	
という		
	本名	、
		は
		：
	通称	は
という	芸名	
		は
という	愛称	
		は
		である
		＝
の	ニックネーム	
		は
		＝
	別名	＝
	Nickname	＝
	Alias	＝

3.4 別名候補の順位付け

Wikipedia ページから別名記述パターンを用いて抽出した別名候補には，対象者の別名のほかに，人の会話部分などノイズが含まれている可能性がある．ノイズの判別を行うため，正式名称と別名候補の関連度を求め，別名候補の順位付けを行う．本研究では，関連度の導出に検索エンジンでのヒット件数に基づいた共起情報を利用することとした．

通常，2 つの語句を組み合わせて AND 検索を行ったとき，そのヒット件数は 2 つの語句が Web 上で共起する頻度を表し，2 語句の関連度を測ることができる[4]．しかし，このヒット件数では，2 語句がどのような文脈で出現

しているのかを知ることができない．そこで本研究では，別名の意味を記載する際に多く用いられる「こと」を正式名称の前に加えて関連度を測ることとする．例えば「松井秀喜」と「ゴジラ」の関連度を測る際は，

“ゴジラ” AND “こと松井秀喜”

をクエリとして検索を行う．これにより，特に別名について書かれた文脈で出現する 2 語句の関連度をヒット件数から測ることができると考えられる．

本研究では，Overlap 係数を用い，別名候補と正式名称の関連度を算出した．また，Google（http://google.com）を用いてそれぞれのヒット件数を取得することとした．

3.5. 人物判別のための別名へのラベル付与

過去に同姓同名人物の識別を目的とした研究が行われており，属性情報をラベルとした人物識別手法が提案されている [4]．本研究では，複数の著名人に該当する別名から特定の人物を判別するため，はてなキーワードのカテゴリに基づいた別名への特徴ベクトルの付与を行う．はてなキーワードとは，Wikipedia と同様，ユーザによって編集が行われるという特徴を有するオンライン百科辞典であり，40 万件以上の単語が登録されている（2013 年 7 月時点）．それぞれの単語には， 20 種類のカテゴリが一つ以上付与されており，本研究では，「読書」「音楽」「映画」「テレビ」「コンピュータ」「食」「アニメ」「ゲーム」「マンガ」「動植物」「ウェブ」「地理」「社会」「アート」「アイドル」「スポーツ」「サイエンス」の 17 種類のカテゴリをラベルに用いる．別名への特徴ベクトルの付与は，対象とする別名の抽出に用いた Wikipedia ページに基づいて行う．まず，Wikipedia ページからはてなキーワードに登録された単語を抽出する．次にそれぞれの単語が属するカテゴリを取得し，カテゴリ頻度を求める．そして求めた各カテゴリ頻度を，別名の特徴ベクトルとし付与する．なお，実際に人物の判別を行う際には，まず識別を行う検索結果のカテゴリ頻度を同様に求め，特徴ベクトルの付与を行う．次に，その特徴ベクトルと各著名人に対応する別名の特徴ベクトルとの相関係数をそれぞれ算出する．最も高い相関値が得られた別名から，その人物について書かれた検索結果であると判定する．

4. 評価実験

日本語版 Wikipedia の 2013 年 5 月 30 日のダンプファイル（http://dumps.wikimedia.org/jawiki/20130530/）から，野球選手や，歌手，政治家など，様々な職種からそれぞれ約 10 名ずつ抽出し，さらに 10 名の外国人著名人を加えた計 153 名に対して提案手法を用いた別名抽出実験を行った．なお，あらかじめ正解である別名は手動で収集し正解データとした．1 人の人物に対して複数の別名が存在するケースがあるため，153 名に対する正解別名の総数は 459 個であった．

また，複数の著名人に該当する別名に対して，提案手法を用いた検索結果の人物判別実験を行う．対象とする検索結果として，今回はニュース記事に限定した．2014 年 1 月 19 日に「あっちゃん」をクエリとして取得した前田敦子さんに関する記事 17 件と中田敦彦さんに関する記事 4 件の計 21 件，及び「リーダー」をクエリとして取得した城島茂さんに関する記事 4 件と大野智さんに関する記事 6 件の計 10 件，合計 31 件を取得した．

4.1. 別名抽出実験の結果と考察

提案手法を用いて別名の抽出を行った結果，合計 506 個の別名候補が抽出された．提案手法による別名候補抽出の精度を評価するため，適合率，再現率，適合率と再現率の調和平均である F 値を算出した（表 2）．

$$\text{適合率} = \frac{C}{A} \tag{2}$$

$$\text{再現率} = \frac{C}{B} \tag{3}$$

ただし，A は提案手法により抽出された別名候補の総数，B は正しい別名の総数，C は提案手法により抽出された別名候補の中で正しい別名の数である．表 2 より，提案手法により著名人の別名が比較的高い精度で抽出できていることがわかる．特に，再現率の値が 0.822 と高かったことから，設定した 22 パターンを用いることで多くの別名が抽出できたと判断できる．また，表 3 に別名抽出結果の例を示す．例えば，前田敦子の別名としては「あっちゃん」「敦子」「まえあつ」「ヘビーローテーション」「不動のセンター」「絶対的エース」「上からマリコ」「チャンスの順番」が抽出されたが，このうち「ヘビーローテーション」「上からマリコ」「チャンスの順番」は AKB48 の楽曲名であり別名ではない．ただし，関連度に基づく別名らしさの順位付けではそれぞれ 4 位，7 位，8 位とヘビーローテーション」以外は下位であることから，関連度を利用することでより適切な別名を抽出することが可能になると考えている．

4.2. 別名からの人物判別実験の結果と考察

別名抽出対象とした 153 名のうち，内田篤人さん，前田敦子さん，中田敦彦さんから「あっちゃん」の別名が，

城島茂さん，大野智さんから「リーダー」の別名が抽出された．以上5名のWikipeidaページから，提案手法を用いて特徴ベクトルを別名ごとに付与し，別名ごとの人物判別実験を行った．実験の結果，「あっちゃん」をクエリとして取得した記事のうち， 21 件中 16 件（76.2%）を正しく判別することができた．一方，「リーダー」をクエリとして取得した記事では，1件以外のすべての記事を城島茂さんに関するものと誤って判別された．これは，城島茂さんと大野智さんが同じアイドルという職種であり，特徴ベクトルの差が少なかったため，うまく判別を行えなかったと考えられる．以上のことから，同じ職種を持つ人物が同一の別名を持つ場合を除いては，提案手法を用いることで，ある程度の人物判別が行えると判断できる．

5. まとめと今後の課題

本研究では，Wikipediaを用いて人物の別名を自動的に抽出する手法と検索結果の人物判別を提案した．Wikipediaは多くのユーザによって，定期的に更新がなされている．そのため，別名のように，次々と新しいものが生成される可能性の高い情報の抽出には有効である．また Wikipedia における特徴的な別名記述パターンを用いることで，ノイズの少ない別名抽出ができたと考える．また，検索エンジンのヒット件数から算出した別名候補と正式名称の関連度に基づく別名らしさの順位付けを行うことにより，抽出してしまった別名以外の文字列を削除できる可能性を示すことができた．さらに，別名が複数の著名人に該当する場合，はてなキーワードのカテゴリを基に作成した特徴ベクトルを用いることで，ある程度の人物判別が行えると考えられる．

今後は，抽出パターンを自動的に更新する方法や，同じ職種を持つ人物が同一の別名を持つ場合の人物判別手法について検討したい．

参考文献

[1] T. Hokama and H. Kitagawa, “Extracting Mnemonic Names of People from the Web”, Proc. of International Conference on Asian Digital Libraries 2006, pp.121-130, 2006.

[2] D. Bollegala, Y. Matsuo, and M. Ishizuka, “Automatic Discovery of Personal Name Aliases from the Web”, IEEE Transactions on Knowledge and Data Engineering, Vol.23, No.6, pp.831-844, 2011.

[3] R. L. Cilibrasi and P. M. B. Vitanyi, “The Google Similarity Distance”, IEEE Transactions on Knowledge and Data Engineering, Vol.19, No.3, pp.370-383, 2007.

[4] X.Wan, J. Gao, M. Li and B. Ding, “Person resolution in person search results: WebHawk”, Proc. of the 14th ACM International Conference on Information and Knowledge Management, pp. 163-170, 2005.

表 2　別名抽出実験結果

適合率	再現率	F 値
0.749 (379/506)	0.822 (379/461)	0.784

表3　別名抽出結果と正解の例

順位	前田敦子	野茂英雄	木梨憲武	渡邉恒雄	アルベルト・ザッケローニ
1	あっちゃん	ドクターK	憲武	独裁者	ザック
2	敦子	トルネード投法	ノリさん	メディア界のドン	ザッキ
3	まえあつ	ドジャータウン	ノリちゃん	政界フィクサー	ビッグ3
4	ヘビーローテーション	The Tornado		俺は最後の独裁者だ	
5	不動のセンター				
6	絶対的エース				
7	上からマリコ				
8	チャンスの順番				
正解	あっちゃん 敦子 まえあつ 不動のセンター 絶対的エース	ドクターK The Tornado	憲武 ノリさん ノリちゃん	独裁者 メディア界のドン 政界フィクサー ナベツネ	ザック ザッキ