人名の別表記の自動抽出手法

古澤秀介，森田和宏　泓田正雄，青江順一
徳島大学　工学部　知能情報工学専攻

インターネットの普及や検索技術の向上により,検索エンジンにキーワードを入力するだけで,インターネットから大量の情報を獲得できるようになった.
検索エンジンでは，キーワード検索が一般的であるが，意味が同じで表記の異なる語がある場合，ユーザー側でキーワードの拡張や検索結果の絞り込みを必要とするという問題がある．これらの問題に対して,現在ではシソーラスを使用して検索支援をする研究が行われている.しかし,人名に対してのシソーラスは存在しない.
そこで本研究は，人名に対してＷｅｂ上からコーパスを収集し，そのコーパス内から人名のニックネームや愛称などの別名を抽出し，人名のシソーラスを構築する手法を提案する．また，実験により提案手法の有効性を実証する．

Automatic selection method for the representation of the name of a person

Shuusuke Furusawa,Kazuhiro Morita,Masao Fuketa, Junichi Aoe
Department of Information Science and Intelligent System, The University of Tokushima

The spread of the Internet and the development of search functions have enabled us to get a large amount of information just by typing a simple keyword.
It is the most common way to use various searching engines, however, if there are several synonyms for the keyword, users have to enter some additional terms or narrow the search results in order to find the specific information they are looking for the name of a person. To deal with these problems, the recent study's offered them a faster way of seraching for required information by means of using a thesaurus. Yet, the thesaurus to the name of a person doesn't exist.
Thus, the following study is not only to show the method of constructing a thesaurus of the name of a person by collecting corpora from the Web sites, and picking up the necessary items such as his nickname or petname but also to prove the availability of this method through experiments.

１章　はじめに

近年，インターネットの普及や検索技術の向上により,検索エンジンにキーワードを入力するだけで,インターネットから大量の情報を獲得できるようになった.それに伴い,効率的な情報収集のためにユーザーの検索システムへのニーズは高まっている.しかし,現在の検索エンジンは,表記による検索をおこなうため,ユーザーが知りたい情報を一度の検索でえることは難しい．　例えば、「ラーメン」を検索キーとして検索しても漢字のラーメン（拉面）に対しての検索結果は得られない.この問題を解決するために，シソーラスを用いて検索を行う支援がなされている[1][2]．しかし人物に対するシソーラスは存在しない.
そこで本研究は,人名に対してＷｅｂ上からコーパスを収集し,そのコーパス内から人名のニックネームや愛称などの別名を抽出し，　人物に対してシソーラスに類似した検索を支援する辞書的なものを構築する手法を提案する．

２章　シソーラス

２．１　　シソーラスとは

シソーラスは一般名詞の意味的用法を表す２７１０個の意味属性（ノード）の上位―下位，関係，全体一部分関係が木構造でしめされたものであると伊藤らは定義している[4]．ノードに属する名詞としては約１３万語がと登録されている．

２．２　　語句同士の関係

シソーラスには語を同義語や反義語や関連語（上位・下位語）などがある.

２．２．１　同義語

同義語とは,同じ意味の語である.たとえば,「私」,「僕」,「我」だけでなく,「特別急行」が「特急」と短縮された語や,「インタフェース」と「インターフェイス」や「犬」と「いぬ」のような表記の揺れも含めた語の関係をいう.

２．２．２　反義語

意味が対立する語である．たとえば,「善，悪」,「売る，買う」のような反対の言葉だけでなく「兄，弟」,「兄，姉」など年齢や性別で対立する語句も反義語に含まれる.

２．２．３　関連語

ある程度の意味的な関連性を持つ語の関係を言います.「肉」と「野菜」は食材という上位語があり，またその逆に「食材」の下位語が「肉」,「野菜」となる.

２．３　　人に対するシソーラス

人に対してもシソーラスは存在する.「～教授」,「～社長」,「～総理大臣」などの役職や肩書きなどをつける表記もその人に対しての同じ意味の語である.また「キムタク」や「柔ちゃん」のようにあだ名や略称なども同じ意味の語である.本研究ではこれらの表記を人名の別表記と定義し,その人名の別表記を以下の２つに分類した.

- 人名を含む人名の別表記

人名を含む実態表記とは,人名の姓もしくは名を含む別表記である.

例えば,「谷亮子」の人名を含む実態表記には,「谷選手」,「谷亮子選手」,「柔道家谷亮子」などがある.

- 人名を含まない人名の別表記

人名を含まない別表記とは，人名が入らない呼称のような表記である.

例えば,「谷亮子」の人名を含まない実態表記とは「ヤワラちゃん」,「ＹＡＷＡＲＡちゃん」,「柔ちゃん」などがある.

３章　人名の別表記の自動抽出

本章では人名の別表記の抽出方法について述べる.人名の別表記は 2 章でも述べたように人名を含む表記と含まない表記に分かれる．2．1 節では人名を含む人名の別表記の抽出方法，2．2 節では人名を含まない実態表記の抽出方法についてそれぞれ述べる.

３．１　人名を含む別表記の自動抽出

本手法は,入力人名に対して入力人名を検索キーとしたコーパス内から，人名の別表記を抽出する．以下に処理の流れを説明する.

Step1:コーパス収集
入力人名を検索キーとして，検索エンジンを使用し，コーパスを取得する.

Step2:人名の別表記候補の抽出
字種切りを使用して，漢字，数字，ローマ字，カタカナのみを残し人名の別表記の候補を抽出する.
Step3:人名の別表記候補の絞り込み
抽出した語句から入力人名を含む語句を取得する.

Step4:形態素解析
取得した語句に対して，形態素解析をおこなう.

Step5:不要語の削除
人名の別表記には不要な形態素を削除する.この不要な形態素を不要語と定義し，不要語を削除する.

不要語の判定には形態素の表記または品詞を用いる.入力人名の前後で削除する不要語は異なるが,削除する方法は，同じで，たとえば前の場合であれば，人名より前に不要語があれば，その不要語も含め，それより前の形態素列を削除する.後の場合は,不要語以降の形態素を削除する.表１，表２に人名の前後で削除される不要語の例を示す.

表１：人名の前での不要語

品詞	語句
時詞	１日,2 日
地名	サイト
数詞	壁紙
接尾語	詳細
接尾助数詞	一覧

表２：人名の後での不要語

品詞	語句
人称代名詞	殿
地名	君
接続詞	様
副詞	氏
名詞	

「中田英寿」を例として人名を含む呼称の抽出を以下に示す.
Step1:「中田英寿」を検索キーとして１００件のコーパスを取得する.
以後，コーパス内の「一際目立ったのが世界選抜の７番中田英寿選手」という文を例にとり人名の別表記の抽出方法を説明していく.

Step2: 字種切りを使って，下記のように区切りができ,人名の別表記候補を取得する.

ボローニャの 16 番**中田英寿選手を取材する**,
ここでは「ボローニャ」,「16 番中田英寿選手」,「取材」を抽出する.

また, 入力が「中田英寿」との文字列比較をおこなう. 結果として「ボローニャ」,「16 番中田英寿選手」,「取材」から「16 番中田英寿選手」のみ抽出する.

Step3: Step２で抽出された「16 番中田英寿選手」に対して形態素解析をおこなう.
図１に「16 番中田英寿選手」の形態素解析結果を示す.

Step4:形態素解析結果を用い, 不要語の削除をおこなう. Step3 で抽出された「16 番中田英寿選手」の場合, 人名の前の形態素「番」の品詞が名詞であるため不要語と判定される, よって,「16」も含めた「16 番」が不要語として削除される. また人名の後の形態素の品詞は「地位名」であり表記も不要語にあてはまらないので削除しない.
結果,「中田英寿選手」が人名を含む人名の別表記として出力される.

16 番中田英寿選手

表記	品詞
１６	数詞
番	名詞
中田	名前（姓）
英寿	名前（名）
選手	地位名

図１　形態素解析結果

３．２　人名を含まない別表記の自動抽出

本手法は, 入力人名に対して「こと＋入力人名」を検索キーとしたコーパス内から, 人名の別表記を抽出する.
人名を含まない実態表記は「やわらちゃんこと谷亮子」のように「○○こと＋人名」の「○○」に表記されていることが多く, 検索キーは,「こと＋人名」としてコーパス収集をした.

Step1:コーパス収集
入力人名に対して「こと＋入力人名」を検索キーとして, 検索エンジンを使用し, コーパスを取得する.

Step2:人名の別表記候補の抽出
Step1 で取得したコーパス内の「こと＋人名」の前の１５文字をすべて取得する. 得られた１５文字の実態表記をデータベースに登録し, 後方一致により, 一致部分の出現頻度を集計する.
Step3:人名の別表記候補の絞り込み
データベースと頻度から人名の別表記の候補を得る.

Step4:不要語の削除
人名の別表記として, 文字単位で不適切な語を削除する.

ここで, step2 で１５文字で区切った訳は, １５文字以上長い人名を含まない別表記が無いと判断したからである.

「谷亮子」を例として人名を含む呼称の抽出を以下に示す.
Step1:「こと谷亮子」を検索キーとして１００件のコーパスを取得する.

以後, コーパス内の文で, 以下の３つの文を例にとり人名を含まない別表記の抽出方法を説明する.

文１:「気まずい空気の中でふとテレビを見るとヤワラちゃんこと谷亮子さんが…」
文２;「アテネオリンピックで金メダルのヤワラちゃんこと谷亮子は…」
文３:「ＯＲＩＸの谷佳知とヤワラちゃんこと谷亮子は…」

Step2: 「こと谷亮子」の前１５文字を抽出
文１からは「ふとテレビを見るとヤワラちゃん」が, また文２からは「ピックで金メダルのヤワラちゃん」が, 文３からは「ＯＲＩＸの谷佳知とヤワラちゃん」のそれぞれことの前１５文字が抽出される.

Step3：Step2 で抽出された３つの文字列で後方から重なっている文字列を比較し人名の別表記候補を絞り込む. また出現頻度も同時に調べる.
まず,「ふとテレビを見るとヤワラちゃん」と「ピックで金メダルのヤワラちゃん」と「ＯＲＩＸの谷佳知とヤワラちゃん」の一致文字列である「ヤワラちゃん」が出現頻度３の実態表記候補となる. また「ふとテレビを見るとヤワラちゃん」と「ＯＲＩＸの谷佳知とヤワラちゃん」の一致文字列である「とヤワラちゃん」も出現頻度２の実態表記候補となる.

Step4：不要語の削除がおこなわれる.
「とヤワラちゃん」は「と」という不要語が存在するので削除する.「と」は２文字目の「ヤ」と字種が異なり, なおかつ「ヤワラちゃん」という別表記候補が存在するので「とヤワラちゃん」は別表記としては不適切と判断する.
結果,「ヤワラちゃん」が出現頻度３として出力される.

４章　実験・考察

３章で述べた手法の評価実験を行った.

４．１　実験

人名の別表記の自動抽出手法は, 入力人名を１００人として, Google 検索によりコーパスを収集し, ３章であげ

た２つの手法に対して結果の抽出をおこなった. 人名の選択はランダムである．正誤判定は人目でおこなった.
正解率は以下のように定義する.

$$\text{正解率} = \frac{\text{正しく抽出できた数}}{\text{抽出した個数}} \ (\%)$$

４．２　実験結果

人名の別表記の自動抽出手法は１００人に対して抽出された人名の別表記は全体で３１１の表記であり, その人名の別表記として誤っている表記は４０個であった。正解率は約８７%であった. 表３にそれぞれの手法の抽出数を表４に人名の別表記の自動抽出結果の例, 表５に人名を含まない自動抽出結果の例を示す.

表３：人名の別表記の自動抽出数

	人名含む 人名別表記	人名含まない 人名別表記
抽出数	２７２	３８
誤抽出	３６	４
正解抽出数	２３６	３４
正解率(%)	８７	８６

表４：人名を含む別表記の自動抽出の結果

入力	呼称	頻度
石井一久	石井一久投手	11
石井一久	石井選手	5
石井一久	石井一久選手	5
中垣内祐一	中垣内監督	8
中垣内祐一	中垣内選手	7
坂本九	大スター坂本九	4
武豊	武豊騎手	45
武豊	武騎手	9

４．３　考察

人名の別表記の自動抽出手法での間違いの表記があった. 一つの例としては歌手の曲名の一部を抽出する誤抽出である.「愛するとは許すこと」という曲を歌う三浦綾子という歌手がいるがこの人のコーパス内には「愛するとは許すこと三浦綾子」という文字列がよくでてくる.

表５：人名を含まない別表記の自動抽出の結果

入力	呼称	頻度
木村祐一	キムキム兄やん	9
木村祐一	キム兄	5
木村祐一	キム	4
木村祐一	キム・ニールヤング	2
加山雄三	大将	35
加山雄三	若大将	33
加山雄三	永遠の若大将	2
河村隆一	ＲＫ	6
河村隆一	ＲＹＵＩＣＨＩ	5
河村隆一	ЯＫ	4
河村隆一	Ｒ・Ｋ	3

すると人名を含まない別表記として「愛するとは許す」という間違った実態表記が抽出される. このような誤抽出に対しての対処方法は検討中である.

５章　まとめと今後の課題

本稿では人物に対してシソーラスに類似した検索を支援する人名の別表記の抽出手法について述べ, ２つの手法は有効であることを確認した.

今後の課題としては, 提案手法の精度向上とともに, 入力人名に対する「読み」の抽出や, 同表記で同じ読みの人物であっても, 職種などで区別することができるような「属性」の抽出により, 人物に対する知識データを作成することである.

[参考文献]
[1] 山本　英子，梅村　恭司　“辞書を用いない関連語リストの構築方法”　「自然言語処理」No.148 － 012, pp81-88, 2001
[2] 塩見　隆一，徳田　克己，青山　昇一，柿ヶ原康二　“シソーラスを用いた文書データの自動分類法”「自然言語処理」No.117 － 014　pp98-pp104,2001
[３] htto://www.gengokk.co.jp
[４] 伊藤 俊介，渡部 広一，河岡 司”情報検索における未知語理解支援方式 ～未知語のシソーラスノードへの分類～ No.159 - 010”pp61-pp66 研究報告 「自然言語処理」