ウェブを用いた外国人名事典の自動編纂

榊原　洋平[†]　　佐藤　理史[†]

[†] 名古屋大学大学院 工学研究科

1. はじめに

英語のテキストを日本語に翻訳する際，そこに現れる外国人の人名は，その発音に基づいてカタカナ表記される (翻字される) のが普通である．しかし，そのカタカナ表記を決定する作業は，次のような理由により，それほど容易ではない．

(1) 英語のテキストには，発音の推測が難しい非英語由来の人名がしばしば現れる．

(2) いくつかの音に対しては，カタカナ表記が複数ある (たとえば,「ヴァ」と「バ」) が，その人名において，どのカタカナ表記が定着しているかは，既訳を調べる以外に方法がない．

(3) 翻訳されたテキストのみを読む読者は，カタカナ表記により人物を同定する．そのため，もし，その人物の人名が既に訳されており，その既訳が定着しているのであれば，その既訳とまったく同一のカタカナ表記を用いる必要がある．

このような理由により，外国人名を翻訳する際，翻訳者は，既訳を探すために外国人名事典を調べることになるが，外国人名事典に収録されている人名には限りがあるため，有名人以外は見つからないことが多い．見つからなかった場合は，調査対象を同一分野の既訳テキストやウェブに広げて，その人名の定着している既訳を探すことになる．

本研究では，このうち，最後の「ウェブを調べる」作業を効率化するために，あらかじめ，ウェブから外国人名事典を自動編纂しておく方法を検討する．ウェブから，大規模かつ高品質な外国人名事典を自動編纂することができれば，この事典を引くだけで，現在ウェブから既訳を探している外国人名の大半のカタカナ表記を決定できるようになると考えられる．

自動編纂する外国人名事典の項目は，次の 3 つの情報から構成されるものとする．

(1) 外国人名 (フルネーム) のアルファベット表記 (a と表記する)

(2) 外国人名 (フルネーム) のカタカナ表記 (k と表記する)

(3) その人物に関して参考となる URL

このうち，本論文では (1) と (2) のみを扱う．すなわち，項目を，人名訳語対 $\langle a, k \rangle$ と同一視する．

人名訳語対 $\langle a, k \rangle$ が満たすべき条件は，次の 5 つである．

(C1) a は，ある人物を指し示す．

(C2) a は，アルファベット表記として正しいスペルである．

(C3) k は，ある人物を指し示す．

(C4) k は，標準的に使われているカタカナ表記である．

(C5) a と k は翻訳関係にある．(同一人物を指し示す．)

本研究の目標は，上記の 5 つの条件を満たす人名訳語対を，精度良く，大量に集めることである．

与えられた人名に対して対訳を推定する研究は，これまで，後藤ら [1] の研究をはじめとして多くの研究があるが，本研究は，人名の収集を含む人名事典の自動編纂全体の自動化を目標としている点が，これまでの研究と大きく異なる．

2. 自動編纂手法の概要

本研究では，次の 3 ステップにより，人名訳語対を収集する．

(1) **カタカナ表記の収集**
コーパスから外国人名と思われるカタカナ表記を抽出する．

(2) **人名訳語対候補の収集**
次の 2 つの方法で人名訳語対の候補を収集する．

(a) カタカナ表記からの訳語対候補の収集
(1) で得られたカタカナ表記のそれぞれに対して，ウェブを用いて対応するアルファベット表記を求め，訳語対の候補を生成する．

(b) アルファベット表記からの訳語対候補の収集
(2a) の過程で得られたアルファベット表記のそれぞれに対して，ウェブを用いて対応するカタカナ表記を求め，訳語対の候補を生成する．

(3) **候補の絞り込み**
得られた訳語対の候補から，信頼できるもののみを選ぶ．

それぞれの詳細を，次節以降で説明する．

3. カタカナ表記の収集

自動編纂の最初のステップでは，外国人名と思われるカタカナ表記をコーパスから抽出する．本研究では，コーパスとして，新聞コーパスとウェブコーパスを 2 種類のコーパスを用い, それぞれ異なる手法を適用して人名候補

を収集する．

3.1 新聞コーパスからの抽出

新聞は，用字や表記などが比較的よく統制されたテキストである．外国人名は，標準的な表記で記述されることが多い．また，人名は,「さん」や「氏」などを直後に伴って記述されるのが普通である．この最後の事実を利用し，以下の３つの条件を満たす文字列を，人名候補として抽出する．

(1) カタカナと「・」と「ー」のみで構成されている
(2) 先頭はカタカナ，末尾はカタカナか「ー」であり，文字列中に「・」を含む
(3) 直後に特定の文字列が出現する

ここで特定の文字列とは，次に示す，人名の直後に出現しやすい文字列をさす．

さん，君，氏，ちゃん，著，主演，医師，容疑者，被告，博士，教授，助教授，の息子，の長男，の次男，の三男，の娘，の長女，の次女，の三女

この方法で新聞コーパスから抽出される人名候補は，非常に高い確率で人名であることが期待できる．

3.2 ウェブコーパスからの抽出

ウェブテキストは，用字，表記，文体などの点で，非常に多様なテキストである．新聞とは異なり，多くの誤りを含むことも，その特徴の一つである．

ウェブでは,「さん」などの敬称を伴わずに人名が現れることが多い．また，掲示板などでは,「さん」などの敬称が，社名やハンドルネームなどの直後に出現することも珍しくない．そのため，新聞コーパスで用いた方法は，人名候補収集法として，あまりうまく機能しない．

このような理由により，ウェブテキストからは，次の方法で人名候補を抽出する．

(1) 次の２つの条件を満たすカタカナ文字列を抽出する．
  (a) カタカナと「・」と「ー」のみで構成されている
  (b) 先頭はカタカナ，末尾はカタカナか「ー」であり，文字列中に「・」を含む
(2) 得られた候補のうち，出現頻度が低いものを破棄する．(実験では，3 未満のものを破棄した)
(3) 残った候補のうち，カタカナ文字列の姓もしくは名に相当する部分が，英辞郎 [2] に人名 (姓または名) として掲載されているものを抽出する．

このなかで，人名であることをチェックしている部分は (3) であるが，その精度はそれほど高くない．すなわち，新聞コーパスから得られる人名候補と比較して，ウェブコーパスから得られる人名候補は，より多くの誤りを含む．

4. カタカナ表記からの訳語対候補の収集

前節の方法で得られたカタカナ表記のそれぞれに対して，ウェブを用いて対応するアルファベット表記を求め，訳語対の候補を生成する．アルファベット表記は，次の3 ステップで求める．

(1) カタカナ表記を検索語として検索エンジンを引き，スニペットを得る．
(2) 得られたスニペット中からアルファベット単語列を抽出する．
(3) 抽出した単語列とカタカナ表記の間の翻字関係をチェックし，良好なもののみを残す．

以下では，上記のそれぞれのステップについて詳しく述べる．

表 1 各文字の翻字対応スコアの例

対応文字列 (ア:ロ)	スコア	対応文字列 (ア:ロ)	スコア	対応文字列 (ア:ロ)	スコア
a:a	50	ae:e	80	c:s	80
a:a=	70	ah:a	90	ch:k	180
a:u	39	ah:a=	110	dg:j	190
a:o	40	ar:a=	110	h:f	90

「=」は長音に対応する

4.1 スニペットの取得

日本語テキストでは，カタカナ表記された外国人名の前後にしばしばアルファベット表記 (原綴) が示される．この事実を利用した訳語抽出は，すでに，Nagata ら [3] によって提案されている．本研究でもこの事実を利用し，アルファベット表記の候補を抽出するためのテキストとして，検索エンジンの出力を用いる．具体的には，カタカナ表記を検索語として Yahoo!☆で日本語のウェブページを検索し，タイトルとスニペットを最大 50 件取得する．これらを合わせたものを，以下ではスニペットと記す．

4.2 アルファベット単語列の抽出

得られたスニペットから，以下の条件を満たすアルファベット単語列をすべて抽出する．

(1) 単語は，アルファベットのみで構成されている．
(2) 各単語の先頭文字は，大文字である．
(3) 2 語以上で構成されている．

4.3 翻字関係のチェック

抽出した単語列のそれぞれに対して，カタカナ表記との翻字関係をチェックする．翻字関係とは，アルファベット表記とカタカナ表記の発音に基づく対応関係のことである．手順を以下に示す．

(1) カタカナ表記をローマ字表記に変換する．
(2) 表 1 のようなアルファベットとローマ字間の翻字関係をスコア付けしたテーブルに基づき，できるだけ対応スコアが高くなるようにアルファベット表記とローマ字表記を対応付ける．
(3) 対応付けができずに残った文字に基づいて，最終的な翻字スコアを決定し，翻字スコアがある閾値 (実験では 0.75) 以上のアルファベット単語列を，翻字関係が良好なアルファベット表記として出力する．

例えば，次の例ではアルファベット側の t とローマ字側の u が対応がつかずに残り，それに基づいて翻字スコ

☆ http://www.yahoo.co.jp

表 2 条件の充足状況

	カタカナ表記から		アルファベット表記から
	新聞	ウェブ	
C1	(△)	(△)	(△)
C2	–	–	–
C3	◎	○	–
C4	○	△	–
C5	△	△	△

アは 0.964 と計算される．

例 Kevin Costner : ケビン・コスナー

k:k e:e v:b i:i n:N ˆ:@ c:k o:o s:s t: :u n:n er:a=

翻字チェックは，明らかに対応する可能性がない単語列の排除には有効に機能する．しかしながら，このチェックにパスしたからといって，かならずしも翻訳関係にあるとは限らない．特に，スペル誤りを含むアルファベット表記のほとんどは，このチェックにパスする．

5. アルファベット表記からの訳語対候補の収集

前節の過程で得られたアルファベット表記のそれぞれに対して，前節の方法をちょうど反対方向に適用し，対応するカタカナ表記を求め，人名の訳語対候補を収集する．使用するスニペットは，アルファベット表記を検索語として日本語ウェブページを検索することによって得られたスニペットであり，そこから，翻字関係を満たすカタカナ表記を抽出する．

この逆方向の適用では，ひとつのアルファベット表記に対して得られる複数のカタカナ表記は，表記の揺れである場合が多い．

6. 候補の絞り込み

最後のステップでは，前節のステップで収集した人名の訳語対候補のうち，信頼できるもののみを選ぶことを行なう．

6.1 条件の充足状況

前節で収集された訳語対が，1 節で示した 5 つの条件をどの程度満たしているかをまとめたものを表 2 に示す．新聞コーパスから抽出したカタカナ表記は，ほぼ人名と考えて良く，その表記も標準的なものと考えて良い．一方，ウェブコーパスから抽出したカタカナ表記は，人名である可能性は高いが，新聞コーパスから抽出したものに比べて確実性は低い．また，ある程度の使用例がある表記ではあるが，標準的かどうかはわからない．C5 の翻訳関係は，まったく可能性がないものは排除されているが，確実性には乏しい．C1 は，C3 と C5 がともに成り立てば成り立つが，十分にチェックされているとはいえない．C2 はまったくチェックされていない．

アルファベット表記から出発して得られた訳語対は，そもそものアルファベット表記自体が，人名であることの確実性に乏しい．

6.2 手法 1: 対応関係の強い訳語対の選択

ここでは，まず，次の条件をみたす訳語対 $\langle a, k\rangle$ を抽出する．

(1) カタカナ表記 k から抽出されたアルファベット表記のうち，スニペット中の頻度が最も高い (単独 1 位の) ものは，a である．

(2) アルファベット表記 a から抽出されたカタカナ表記のうち，スニペット中の頻度が最も高い (単独 1 位の) ものは，k である．

つまり，k から探した場合に a が最も有力な候補であり，a から探しても k が最も有力な候補である場合，訳語対 $\langle a, k\rangle$ を，信頼できる訳語対として抽出するということである．これは，C5 を両方向からチェックすることに相当し，かつ，C1 を C3&C5 によって間接的にチェックすることに相当する．

すでに述べたように，カタカナ表記から探した場合は，スペルが誤っているアルファベット表記が候補に残る．一方，アルファベット表記から探した場合は，非標準的なカタカナ表記 (表記のゆれ) が候補に残る．「正しいもの，標準的なものは，最もよく現れる」という仮定をおけば，上記のような両方向で単独 1 位となるペアは，スペルが正しく (C2)，カタカナ表記として標準的である (C4) ことが期待されうる．

6.3 手法 2: 第二の選択

次に，残された訳語対候補 $\langle a, k\rangle$ に，いくつかのテストを適用し，残ったものを信頼できる訳語対として出力する．

(1) スペルエラー等の排除
前節で抽出された訳語対を $\langle a, k\rangle$ とするとき，$\langle x, k\rangle$ が候補に含まれるような x はスペルエラーである可能性が高い．そこで，候補から $\langle x, *\rangle$ を除去する．同様に，$\langle a, y\rangle$ が候補に含まれるような y は表記の揺れである可能性が高いため，候補から $\langle *, y\rangle$ を除去する．

(2) **アルファベット側の人名（C1）チェック**
アルファベット表記 a の姓あるいは名に相当する部分のすくなくともどちらか一方が，英辞郎に人名 (姓または名) として載っている場合は残し，載っていない場合は破棄する．

(3) **アルファベット側のスペル（C2）チェック**
カタカナ表記候補 k に対して複数の訳語対候補 $\langle a_i, k\rangle$ が存在する場合，a_i のヒット数が最も多いものを残し，それ以外は破棄する．これは，スペルエラーなどの誤った表記よりも正しい表記の方がウェブ上には多く存在するという仮定に基づく．

(4) **対訳関係（C5）チェック** 1
a と k のアンド検索のヒット数がある閾値 (実験では 5) 未満の項目を破棄する．

(5) **対訳関係（C5）チェック** 2
アルファベット表記 a に対して複数の訳語対候補 $\langle a, k_i\rangle$ が存在する場合，a と k_i のアンド検索のヒット数が最も多いものを残し，それ以外は破棄

表 3 実験結果

	新聞	ウェブ
抽出されたカタカナ表記	37,925	90,140
収集された訳語対候補 (順方向)	12,416	63,516
カタカナ表記の異なり	10,041	50,294
アルファベット表記の異なり	12,010	57,586
収集された訳語対候補 (逆方向)	18,392	84,284
カタカナ表記の異なり	16,209	69,638
アルファベット表記の異なり	10,807	51,840
収集された訳語対候補 (合計)	20,978	101,009
カタカナ表記の異なり	17,859	79,623
アルファベット表記の異なり	12,010	57,586
手法 1 で抽出された訳語対	6,922	29,988
手法 2 で抽出された訳語対	1,727	7,623
訳語対の総計	42,239	

表 4 検証結果

		正しい	誤り
新聞	手法 1	98	2
	手法 2	96	4
ウェブ	手法 1	88	12
	手法 2	70	30
新聞＋ウェブ		87	13

する．これによりカタカナ表記とアルファベット表記の対応関係が強い項目を選ぶことができる．但し，k_i が k_j の部分文字列となる場合 (例えば k_i=「バーバラ・カー」，k_j =「バーバラ・カーレイ」) は，翻字スコアが低い方を破棄する．

7. 実　　験

これまで述べてきた方法の有効性を検証する実験を行なった．カタカナ表記候補の収集に用いたコーパスは以下の通りである．

(1) 新聞コーパス：毎日新聞 15 年分 (1995 年–2004 年)
(2) ウェブコーパス：河原ら [4] が収集した「ウェブ上の 5 億文の日本語テキスト」

実験結果を表 3 に示す．新聞コーパスからは手法 1 で 6,922 個，手法 2 で 1,727 個，ウェブコーパスからは手法 1 で 29,988 個，手法 2 で 7,623 個の人名訳語対が抽出された．すべてを合わせた場合の異なりは 42,239 個であった．

これら抽出された人名訳語対の精度をサンプル調査した結果を表 4 に示す．それぞれ，100 個の訳語対を無作為に抽出し，人手で正しいかどうかの判定を行なった．1 節で示した 5 つの条件を全て満たす場合に「正しい」と判定した．この表に示すように，新聞コーパスを用いた場合の精度は良好であったが，ウェブコーパスを用いた場合の精度は，それほど高くなかった．

抽出された訳語対のうち，正しい訳語対の例を表 5 に，誤った訳語対を表 6 に示す．新聞コーパスを用いた場合には，カタカナ表記の抽出誤りやスペルエラーが見られた．たとえば，カタカナ表記の抽出では，「＝」を含めていないため，「ジャン＝ピエール・ランタン」から，「ピエール・ランタン」のみが抽出されていた．一方，ウェブコーパスを用いた場合には，「人名である」という条件を満たさないものが誤りの大半を占めた．これは，カタカナ表記の抽出の段階での人名チェックが不十分であることに起因すると考えられる．

表 5 正しい項目の例

アルファベット表記	カタカナ表記
Adam Kane	アダム・ケイン
Allen Drury	アレン・ドルーリー
Daniel Cohen	ダニエル・コーエン
Giuseppe Sinopoli	ジュゼッペ・シノーポリ
Steven Johnson	スティーブン・ジョンソン

表 6 誤り項目の例

	アルファベット表記	カタカナ表記
新聞	Nigel Fisger (正しくは Nigel Fisher)	ナイジェル・フィッシャー
	Pierre Lentin (正しくは ジャン＝ピエール・ランタン)	ピエール・ランタン
ウェブ	Black Hills	ブラック・ヒルズ (人名ではない)
	Cold War	コールド・ウォー (人名ではない)

8. おわりに

新聞コーパス，ウェブコーパス，ウェブの検索エンジンを利用して，外国人名事典の項目となる人名訳語対を自動的に収集する手法を提案した．新聞コーパスを用いた場合は，およそ 96% の正確さで 8,649 個，ウェブコーパスも用いた場合は，およそ 87% の正確さで 42,239 個の人名訳語対を収集することができた．

我々の当面の目標は，99% の精度を有した収録人名数 3 万件の外国人名事典を自動編纂することである．すでに，正しい訳語対を 3 万件以上の収集できており，件数の目標は達成できているが，精度は全く不十分である．今後，より多くの検証手法を実現し，精度の向上を図る予定である．

参 考 文 献

1) 後藤功男, 加藤直人, 田中英輝, 江原暉将, 浦谷則好: World Wide Web を用いた外国人名の英訳自動獲得, 情報処理学会論文誌, Vol. 47, No. 3, pp. 968–979 (2006).
2) Electronic Dictionary Project: 英辞郎 第二版, ALC (2005).
3) Nagata, M., Saito, T. and Suzuki, K.: Using the Web as a Bilingual Dictionary, *Proc. of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, pp. 95–102 (2001).
4) 河原大輔, 黒橋禎夫: 高性能計算環境を用いた Web からの大規模格フレーム構築, 情報処理学会研究報告, NL-171-12, pp. 67–73 (2006).