外国人名対訳辞典の大規模化 –15 万件の自動編纂–

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対訳辞典を調べることになるが，外国人名対訳辞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対訳辞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対訳辞典を自動編纂することができれば，この対訳辞典を引くだけで，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対訳辞典のエントリは，次の 2 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)

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，後藤ら [1] を初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対訳辞典の編纂全体の自動化を行う点で大きく異なる．我々は昨年におよそ 4 万 3 千件の訳語対を 83% の精度で集めることに成功したが [2]，規模，精度ともに不十分である．本研究では昨年の手法を拡張し，１５万件を越える訳語対を自動的に獲得する．

2. 自動編纂手法の概要

本研究では，次の 2 ステップで人名訳語対を収集する．

(1) **カタカナ表記の収集**
コーパスやウェブから外国人名と思われるカタカナ表記を抽出する．
(2) **エントリの作成**
(a) カタカナ表記からの訳語対候補の収集
(b) アルファベット表記からの訳語対候補の収集
(c) 新たなカタカナ表記からの訳語対候補の収集
(d) 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

自動編集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やウェブから抽出する．本研究では，新聞コーパスとウェブコーパスの 2 種類のコーパスを用い，それぞれに異なる手法を適用して人名候補を収集する．

すべての手法に共通して，次の 3 つの条件を人名カタカナ表記の表記上の条件として利用する．

(1) カタカナと「・」と「ー」と大文字アルファベットで構成されている
(2) 先頭はカタカナもしくはアルファベット，末尾はカタカナとアルファベットと「ー」のいずれかであり，文字列中にカタカナと「・」を含む
(3) 「・」で区切った構成要素がアルファベットを含む場合，その構成要素は一文字である

3.1 手法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 2 つの条件を満たす文字列を，人名候補として抽出する．

(1) 表記上の条件を満たす
(2) 直後に「さん」や「氏」のような人名の直後に出現しやすい文字列が出現する

こうして抽出された人名候補は，非常に高い確率で人名であることが期待できる．

3.2 手法２:新聞コーパスと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，掲示版などでは,「さん」などの敬称が，社名やハンドルネームなどの直後にも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) 表記上の条件を満たす文字列を抽出する
(2) 得られた候補の「人名らしさ」をチェックし，良好なもののみを残す

また，新聞コーパスからも同様の手法で人名が得られると考えられるので，新聞コーパスについてもこれを行う．

人名らしさについては 3.4 節で述べる．

3.3 手法３：ウェブ検索エンジンを用いての収集

新聞コーパスやウェブコーパスはある時点で作成されたものであり，それ以降の情報は含まれない．そこで，より新しい情報源としてウェブを用いる．人名を含むテキストには，他の人名もしばしば出現する．この事実を利用し，カタカナ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（スニペット）を用いる．

具体的にはウェブから次の方法で人名候補を収集する．

(1) 人名カタカナ表記の集合 K を用意する
(2) 集合 $K_q = K$ とする
(3) 集合 $K'_q = \phi$ とする
(4) すべての $k \in K_q$ に対し，(4a) から (4d) を行う
  (a) k を検索語として Yahoo!☆で日本語ウェブページを検索し，タイトルとスニペットを得る
  (b) 表記上の条件を満たす文字列を抜き出す
  (c) 得られた候補の「人名らしさ」をチェックし，良好なもののみを残す
  (d) 残った候補のうち K に含まれていないものを K と K'_q に加える
(5) K_q を K'_q に置き換えて (3) 以降を繰り返す
(6) K を出力する

☆ http://www.yahoo.co.jp

3.4 人名らしさによるフィルタリング

ある語句が与えられたとき，それが人名らしいかどうかを推定する．語句の人名らしさはその語句を構成する要素の人名らしさと構造から定まる，と仮定する．本研究では，構成要素としてファーストネームとラストネームの二つを考慮する．人名らしさを求めるために，まず，語 v のある条件 A らしさを計る指標 $score_A(v)$ を定める．次にこれを用いて語句のファーストネームに相当する部分 v_f のファーストネームらしさ $score_F(v_f)$，ラストネームに相当する部分 v_l のラストネームらしさ $score_L(v_l)$ を求める．最後にこれらを組み合わせることで語句全体の人名らしさとする．

3.4.1 A ら し さ

リスト X とその要素 x に対して定義される条件 A がある．A を満たす要素のリストを $satisfy(A, X)$ と表す．このとき，A の濃度を

$$density(A, X) = \frac{|satisfy(A, X)|}{|X|}$$

と定義する．

いま，次の条件を満たす二つのリスト X_D と X_S がある場合を考える．

- X_D は，X_S に完全に含まれる
- $density(A, X_D) > density(A, X_S)$

ある語 v に対し，それぞれのリストに対する出現確率の差を

$$\delta(v, X_D, X_S) = p(v, X_D) - p(v, X_S) \quad (1)$$

とする．ここで $\delta(v) > 0$ の場合，v は X_D の方によく出現するので「A らしい」といえる．一方 $\delta(v) < 0$ の場合，v は X_S の方によく出現するので「A らしくない」といえる．この出現確率の差 $\delta(v)$ を「A らしさ」を計る指標として用いる．

3.4.2 各要素の人名らしさ

式 1 を用いて，語のファーストネームらしさ，ラストネームらしさを求める．3.1 節で得られた，人名カタカナ表記候補は人名である確率が非常に高い．そこで，このリストのファーストネーム・ラストネームに相当する部分のリストをそれぞれ X_{D_F}, X_{D_L} とする．次に，新聞コーパスとウェブコーパスから，表記上の条件のみでカタカナ文字列を抜き出す．ここには人名以外のカタカナ文字列も多く含まれていると考えられる．このリストのファーストネーム・ラストネームに相当する部分のリストをそれぞれ X_{S_F}, X_{S_L} とする．これらのリストを用い，ファーストネームらしさ，ラストネームらしさを次式で求める．

$$score_F(v) = \delta(v, X_{D_F}, X_{S_F})$$
$$score_L(v) = \delta(v, X_{D_L}, X_{S_L})$$

3.4.3 語句全体の人名らしさ

各要素の人名らしさが求まったので，それらを組み合わせて人名らしさとする．

$$score(\langle v_f, v_l \rangle) = \frac{score_F(v_f) + score_L(v_l)}{2}$$

本研究では $score(\langle v_f, v_l \rangle) > 0$ の場合は人名であると判定し，$score(\langle v_f, v_l \rangle) \leq 0$ の場合は人名ではないと判定する．

4. 収録エントリの作成

4.1 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

こうして得られた単語列のそれぞれと，カタカナ表記とを対にしたものを，人名の訳語対候補とする．

以下では，それぞれのステップについて詳しく述べる．

4.1.1 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[3] ら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 (スニペット) 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 100 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．

4.1.2 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，アルファベットと “.” のみで構成されている
(2) 単語列の末尾はアルファベットである
(3) 各単語の先頭文字は，大文字である
ただし，ミドルネームに相当する部分においては “von” や “de” などを許容する
(4) 構成されている語数が，カタカナ表記を「・」で区切った数と同じか 1 つ多い数である
(5) “.” を含む語は，大文字一文字＋ “.” という構成である

4.1.3 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，カタカナ表記との間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．翻字チェックには従来と同じ手法 [2] を用いる．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，スペル誤りのほとんどは，このチェックにパスする．

4.2 アルファベット表記からの訳語対候補の収集

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，新たな人名の訳語対候補を収集する．このとき，アルファベット表記は前節で得られたアルファベット表記のうち，スニペット中での頻度が最も多かったもののみを使用する．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，表記のゆれである場合が多い．

4.3 新たなカタカナ表記からの訳語対候補の収集

前節で新たに得られたアルファベット表記のそれぞれに対して，4.1 節の手法を適用し，対応するアルファベット表記を求める．

この収集の目的は，新たなアルファベット表記を集めることではなく，のちに行う候補の絞り込みの準備である．

4.4 候補の絞り込み

最後のステップでは，前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．

4.4.1 翻訳関係の強い訳語対の選択

翻訳関係の強い訳語対 $\langle a, k \rangle$ を，次の条件で抽出する．

(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．

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，最も良く現れる」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，カタカナ表記として標準的であることが期待できる．

表 1 カタカナ表記の収集結果

	抽出されたカタカナ表記
手法 1 新聞コーパス	40,035
手法 2 新聞コーパス＋ウェブコーパス	698,471
手法 3 ウェブ検索エンジン	895,005
総計（重複除く）	1,262,723

表 2 エントリ作成結果

	setA	setB	setC
用いたカタカナ表記	40,035	200,000	400,000
代表表記	11,120	74,137	105,985
カタカナ異表記	3,442	19,991	22,425
合計	14,562	94,128	128,410
代表表記の総計		191,242	

表 3 検証結果

	エントリ数	正しい	誤り	正解率
setA	11,120	97	3	97%
setB	74,137	87	13	87%
setC	105,985	94	6	94%
setA + setB + setC	191,242	90	10	90%

4.4.2 カタカナ異表記の付加

既訳が定着していない人名では，標準的なカタカナ表記以外の表記も用いられることがある．本研究では訳語対 $\langle a, k\rangle$ に対するカタカナ異表記 k' を次の条件で抽出する．

(1) 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ものは，a である

(2) k' のウェブ検索単独ヒット数が k のそれの 10 分の 1 以上である

5. 実　　験

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．

(1) 新聞コーパス：
毎日新聞 15 年分 (1991 年–2005 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」

カタカナ表記候補の収集の結果を表 1 に示す．手法 3 では２０万回ほど検索を行った時点でプログラムを停止させた．抽出されたカタカナ表記はすべてを実験に用いるには多すぎるので次のように分類した．

setA　手法 1 で得たカタカナ表記

setB　手法 2 で得たカタカナ表記から setA との重複を除いたもののうち頻度上位２０万件

setC　手法 3 で得たカタカナ表記から setA，setB との重複を除いたもののうち頻度上位４０万件

これらのカタカナ表記に対し 4 節の手法を適用してエントリを作成した．結果を表 2 に示す．

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 3 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った．1 節で示した 5 つの条件をすべて満たす場合に「正しい」と判定した．setA の精度は良好であったが，setB，setC 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 4 に，誤った訳語対の例を表 5 に示す．setA 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．例えば，カタカナ表記の抽出では「＝」を含めていないため,「ジャン＝ベルトラン・アリスティド」から「ベルトラン・アリスティド」のみが抽出されていた．一方，setB，setC では「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 4 正しい項目の例

アルファベット表記	カタカナ表記
Adam Benjamin	アダム・ベンジャミン
Thomas Carlyle	トーマス・カーライル
Nick Broomfield	ニック・ブルームフィールド

表 5 誤り項目の例

	アルファベット表記	カタカナ表記
setA	Ron Sommmer	ロン・ゾンマー（正しくは Ron Sommer）
	Bertrand Aristide	ベルトラン・アリスティド（正しくは ジャン＝ベルトラン・アリスティド）
setB	Pure Rose	ピュア・ローズ（人名ではない）
	Dzanga Sangha	ザンガ・サンガ（人名ではない）
setC	Julie Marie	ジュリー・マリー（人名ではない）
	Knick Knacks	ニック・ナックス（人名ではない）

6. おわりに

新聞コーパスやウェブコーパスを基盤としてウェブから情報を集め，外国人名対訳辞典を従来手法よりも大規模に作成する手法を提案した．提案した手法により人名訳語対をおよそ 90% の正確さで 191,242 項目収集することができた．今後の課題としては，正確さの向上とその人物に関する内容の付加が挙げられる．

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) 榊原洋平, 佐藤理史: 外国人名事典の自動編纂, 言語処理学会第 13 回年次大会発表論文集, pp. 879–882 (2007).

3) Nagata, M., Saito, T. and Suzuki, K.: Using the Web as a Bilingual Dictionary, *Proceedings 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, Vol. NL-171-12, pp. 67–73 (2006).