レファレンス・ツールとしての Web 日英人名検索システム

辻慶太 * 影浦峡 †

Abstract

翻訳者支援を目的として，英語人名に対する日本語訳語を，Web 情報に基づいて出力するシステムを提案する。本システムは，英語人名が入力されると，(1) 対訳辞書が挙げる日本語人名，(2) 翻字パターンから生成されるカタカナ文字列，の集合を Google に送り，ヒット件数の多いものから順に出力する。ヒット件数を，表記の一般性を示す尺度として利用することで，人名に関する「定訳」が出力できると考えている。67 個の英語人名に対して，その 7 割以上に，適切な日本語訳語が出力できることを確認した。

1 はじめに

近年，世界規模の Web の発達を受けて，言語の枠を越えた情報流通が盛んになっている。その中には，Web に新たに発表された各種文書を翻訳し，自身のホームページで公開するといったボランティアベースの報知活動も含まれる。本研究では，そうした翻訳者支援の一環として，英語人名に対する日本語定訳を出力するシステムを提案したい。

本研究のシステム（以下 “QRLex-P”）は，英語人名が入力されると，(1) 対訳辞書が挙げる日本語人名，(2) 翻字パターンから生成されるカタカナ文字列，の集合を Google に送り，ヒット件数の多い順に出力するものである。ヒット件数を，表記の一般性を示す尺度として利用することで，人名に関する「定訳」が出力できると考えている。QRLex-P の利用者には先述のように，新規情報を含んだ文書の翻訳者を想定している。従って QRLex-P は，既存の辞書にはまだ収載されていない新しい人名訳語を出力できることが求められる。(1) だけでなく (2) の翻字も組み込んでいるのはその為である。自然言語処理分野では，多言語処理に翻字を取り入れる研究が増えており，様々な言語の対でその有効性が示されている（Knight & Graehl (1998)，Kang & Kim (2000)，AbdulJaleel & Larkey (2003)）。また Google のようなサーチエンジンは，基本的に商用ベースであり，利用者確保に向けて，常にデータを最新に保っている。新しい人名や「定訳」といった言葉の動的な側面は，迅速に更新されるデータに反映されていると期待できる。

辞書の電子化が進み，また Britanica Online（http://www.britannica.com/）のように Web で利用できる百科事典が増えるなど，近年ではレファレンスツールの電子化・オンライン化が進んでいる。本研究はそうした流れの中で，Web での利用を想定したレファレンスツールを提供するものと位置付けることができる。また，従来の印刷物を越えて，様々な情報が電子化されるにつれ，これまでは図書の書誌事項を提示して終わるなどの形でしか対応できなかった，様々な情報要求に対して，個別に対応することが技術的・環境的に可能になりつつある。本研究はそうした中で翻訳者に対象をしぼり，彼らのニーズに応えるシステムの開発を目指したものと位置付けることができる。

*国立情報学研究所 keita@nii.ac.jp
†国立情報学研究所 kyo@nii.ac.jp

2 QRLex-P の動作仕様

QRLex-P は以下の流れで日本語訳語を出力する。

(1) まず英語人名が入力されたら，対訳辞書を調べる。対応する日本語訳語があったら，それらは候補語 1 とする。

(2) 次に英語人名を以下の「子音字」「母音字」の組合せによって定義される「単位」に分割する。

(2-1) 子音字の定義

- bb, ch, dd, ff, gg, ll, mm, nn, pp, ph, rh, rr, sh, ss, th, tt は 1 つの子音字とみなす。
- a, i, u, e, o 以外の 21 字で，上記以外の形で現れているものも子音字とする。

(2-2) 母音字の定義

- [a,i,u,e,o] × [a,i,u,e,o] gh は 1 つの母音字とみなす。
- aar，eau，oar は 1 つの母音字とみなす。
- 上記以外で aa，au，oa，oo，eu，ee は 1 つの母音字とみなす。
- ar，ar，ir，ir，ur，ur，er，er，or，or は a，i，u，e，o が後接する場合以外は 1 つの母音字とみなす。
- ea は si が後接する場合以外 1 つの母音字とみなす。
- ign では ig を 1 つの母音字とみなす。
- 上記以外の形で現れている a，i，u，e，o は 1 つの母音字とみなす。
- y は a，i，u，e，o が後接していない場合は 1 つの母音字とみなす。

(2-3) 単位の定義

- 子音字＋母音字は１単位とみなす。
- 子音字＋子音字となっている場合，前の子音字は１単位とみなす。
- 語頭の母音字は１単位とみなす。
- 語尾の子音字は１単位とみなす。
- sio，tio，dhi，tsu は１単位とみなす。

(3) 対訳辞書から抽出しておいた翻字パターンを各英語単位に適用し，日本語文字列を生成する。例えば “Bush”は “Bu”，“sh”に分割し，[ブ，バ...] × [シュ，シ...] のような文字列を生成する。これらは候補語 2 とする。

(4) 辞書引きによる候補語 1 と，翻字パターンによる候補語 2 を Google に入力し，ヒット件数順に表示する。[1] その際，入力された語と候補語の AND 検索と，候補語のみの単独検索の２通りを行う。前者は，元の英語と共に現れているページ数が多いほど，その日本語は訳である可能性が高いという仮説に基づいている。前者の AND 検索は精度を追求し，後者の単独検索は再現率を追求しているとも言える。

翻字パターンは，対訳辞書に含まれる日英対を，英語に関しては上記と同じ基準で単位分割し，日本語に関しては基本的にカタカナ１文字を１単位（ただし「ァ」「ィ」「ゥ」「ェ」「ォ」「ッ」「ャ」「ュ」「ョ」「ー」だけは直前の文字と一緒にして１単位）として単位分割し，日英対における単位同士の対応可能性を Melamed (2000) の訳語対抽出手法で推定して獲得したものである。

(3) では単位同士の対応可能性が高いものから順に，翻字パターンとして適用している。また組合せ爆発を避ける為，適用するパターンは対応可能性が高いものに限定した。具体的に言うと，1，2，3，4 単位から成る英語人名に対しては，単位ごとにそれぞれ 64，8，4，3 個の翻字パターンを適用した。

なお QRLex-P では，姓名は個別に検索されることを想定している。ある個人を指した姓名の同時検索は今後の課題としたい。また以下では便宜上，適切な訳語をヒット件数に基づいて上位に出力することを “validation”と呼ぶことにする。

3　辞書と調査対象人名

以下ではまず本研究が用いた対訳辞書と，実験・調査の対象とした人名について述べる。

3.1　対訳辞書

本研究では対訳辞書として，日外アソシエーツの『カタカナから引く外国人名綴り方字典』（以下「日外」）と EDP の『100 万語収録のスーパー英和・和英辞典：英辞郎』（以下「英辞郎」）の２つを用いた。前者には 112,679 個の日英人名対が含まれていた。後者には約 140 万個の訳語対が含まれていたが，そこからカタカナ語と英語の対 149,134 個を抽出して用いた。両者を合わせて重複を除いたところ，229,029 個の対が得られ，これらを QRLex-P の辞書対とした。先述の単位分割方法及び Melamed (2000) の手法によって抽出された翻字パターンの数は 8,787 であった。

[1] 各ページの信頼性などは考慮しない。どのページもいわば同じ１票とみなしている。

3.2　調査対象人名

本研究は，翻訳者を支援し，従来の辞書にない機能を提供するシステムの構築を目指すものである。そこで本研究では，(1) 辞書にない人名が，翻字と validation によってどれだけ提供できるか，(2) 辞書にある人名については「定訳」が，validation によってどれだけ提供できるか，を検証することにしたい。(1) に関して今回は，最近実際に翻訳が行われた英語ページに含まれる人名（これらの中には辞書にないものも含まれる）を検証対象とする。(2) の「定訳」は把握が難しいのだが，今回は上記英語ページの日本語版で使われている日本語人名を中心に据え，それらを適宜筆者らの判断で修正したものを定訳とした。

本研究では，政治・社会及びコンピュータに関する英語記事における人名を対象とした。まず *Counter Punch* の記事 *Our Heroic Baby Killers*（2004/9/25），*Against a Countrywide Rebellion, the Capture of Samarra is a Bloody, But Useless, Gesture*（2004/10/6）に現れていた 14 人名を抽出した。これらの記事はいずれも益岡・いけだらのブログ（http://humphrey.blogtribe.org/）で日本語訳が提供されている。

次に *Common Dreams*，*New York Daily News*，*Center for American Progress* の記事それぞれ *Boom Time for Billionaires*（2004/10/05），*The War's Littlest Victim*（2004/09/29），*Afghanistan in Crisis ? Facts and Figures*（2004/10/4）に現れていた 31 人名を抽出した。これらの記事はいずれも「暗いニュースリンク」（http://hiddennews.cocolog-nifty.com）で日本語訳が提供されている。

上記は政治・社会に関する記事における人名である。分野によって傾向が変わることも考えて，以下のコンピュータ関連記事における人名も取り上げた。即ち，internetnews.com の記事 *MS, Intel Shepherd New Web Services Spec*（2004/10/8），*FTC Pursues Former Spam King in Court*（2004/10/8），*Transmeta Unveils New Efficeon*（2004/10/7），*IBM Mainframes Outfitted For On Demand*（2004/10/7），*Kodak, Sun Settle Patent Case*（2004/10/7），*Conway Gets His Chance To Defend PeopleSoft*（2004/10/6）に現れていた 24 人名である。これらの記事はいずれも IT media（http://www.itmedia.co.jp/）で日本語訳が提供されている。

本研究では，上記の 14，31，24 人名から重複を除いた 67 の人名を調査対象とした。

4 結果の概要

調査対象の 67 個の英語人名のうち，QRLex-P による AND 検索あるいは単独検索で，定訳がトップに出力されたのは 51 個であった。即ち，51/67=76.1 %の割合で，定訳の出力に成功した。以下で内訳を見ていく。

まず日外に何らかの日本語訳語が載っていた語は 44 個で，そのうち定訳が載っていた語は 12 個であった。validation による出力では：

- AND 検索と単独検索の両方でトップ...10
- AND 検索のみトップ...1
- 単独検索のみトップ...0
- トップにならなかった...1

一方，英辞郎に何らかの日本語訳語が載っていた語は 51 個で，そのうち定訳が載っていた語は 42 個であった。validation による出力では：

- AND 検索と単独検索の両方でトップ...27
- AND 検索のみトップ...11
- 単独検索のみトップ...0
- トップにならなかった...4

上記のように AND 検索による成績は良好で，単独検索に比べて「再現率」は犠牲にならないことが分かった。

訳語数	0	1	2	3	4	5+
日外	23	13	7	7	3	14
英辞郎	16	25	9	8	2	7

表 1: 辞書に挙げられていた訳語数

調査対象の 67 人名に関して，日外・英辞郎に挙げられていた訳語数を調べたところ表 1 のようになった。訳語が挙げられていた場合，その平均個数は日外では 4.9 個，英辞郎では 2.3 個となっている。[2] 即ち，重複を除いても，一人名当たりおよそ 5，6 個の訳語が validation にかけられることになる。そうした中，上記の度合いで定訳がトップに出力されるということは，validation が適切に機能したと考えることができる。

調査対象の 67 人名のうち，日外・英辞郎に定訳が挙げられていなかったものは 18 個であった。これらはいわば翻字に頼るしかない人名である。これら 18 人名のうち，翻字によって定訳が生成され，かつ AND 検索・単独検索いずれかでトップに出力されたものは 6 人名であった（AND 検索と単独検索の両方でトップは 4 例，単独検索のみトップは 2 例）。即ち，翻字による定訳生成及び validation 全体の成功率は約 33 %であった。

[2] 日外は「カタカナから英語人名を引く」のが目的なので，英辞郎に比べてカタカナ表記は多様になっている。

5 個別分析

辞書に定訳が挙げられていながら validation でトップに出力されなかった人名は以下の通りである。サンプルは少ないが，他の辞書訳語が上位に来るケースが多かった。

- Wallace：ウォリス（他の辞書訳語が上位）
- Juan：ホアン（同上）
- Gerard：ジェラルド（同上）
- Fersht：ファーシュト（翻字によって生成された「ファヒト」などの無関係な語が上位）

さて辞書に定訳が挙げられていなかった先ほどの 18 人名のうち，翻字で適切に処理できた 6 人名は以下の通りである：

- Allawi：アラウィ
- Iyad：イヤド
- Goldby：ゴールドビ
- David：デビッド
- Mastrobattista：マストロバチスタ
- Lindorff：リンドフ

次に，辞書に定訳が挙げられておらず，かつ翻字・validation によって適切に処理できなかった 12 人名に関する失敗のタイプと原因・対策を挙げる：

タイプ 1：訳語の生成に失敗

(1) Boylan：ボイラン（“Bo”に「ボー」「ブー」などが優先的に適用された）
(2) Muqtada：ムクタダ（“Mu”に「マ」「ミュ」などが優先的に適用）
(3) Durakovic:ドラコビッチ（“Du”に「ダ」「デュ」などが優先的に適用）

これらは，対応可能性が低い翻字パターンも広く適用することで避けることができる。

(4) Dietz：ディエズ（“tz”を 2 単位に分解）
(5) Khalilzad：カリルザッド（“Kha”を 2 単位に分解）
(6) Zalmay：ザルメイ（“may”を “ma”と “y”に分解し，前者に「マ」等を適用していた）
(7) Axel：アクセル（“xe”は日本語の 2 単位「クセ」に対応していた）
(8) Buffett：バフェット（“ffe”に対する「フェッ」の優先順位が低い）

これらについては，英語の単位の取り方を変えることで対応できる可能性がある。例えば “tz”や “kha”は 1 単位とするなどである。また日本語については，基本的に 1 文字 1 単位としたが，“x”に対しては特別に扱うことも考えられる。さらに「ッ」は前の単位に付けた形で 1 単位としていたが，後の単位に付けた方が英語との対応が取りやすいように思われる（上記の例で言うと，“ffe”と「フェッ」ではなく，“tte”と「ット」を対応させる）。

(9) Gonzalez：ゴンザレス（最後の “z”には「ズ」や「ツ」などが優先的に適用された）
(10) Enderle：エンダール（“le”に「レッ」「リー」などが優先的に適用）
(11) Telesca：テレスカ（同上）

語中の単位と語尾の単位とでは，翻字パターンの傾向が異なる可能性がある。例えば “le”は語尾では “ル”が多いなどである。単位の位置を考慮することで，これらの誤りを防げるかもしれない。

タイプ 2：validation で失敗

(12) Dave：デーブ（「デイヴ」「デイヴィッド」などが上位に出力）

これには上記の「デイヴィッド」のように正式名や愛称が絡んでくる場合があった。このような誤りは，辞書を別途用いて処理することで改善できる可能性がある。

さて辞書に定訳が挙げられていた／いなかった英語人名，後者に関してはさらに，翻字・validation で適切に処理できた／できなかった英語人名，の全体傾向を把握する目的で，それらの単位数と定訳のヒット件数を調べてみた。

まず英語人名の単位数は表 2 のようになった。表 2 で「辞書あり」は辞書に定訳が挙げられていた英語人名を指す。「翻字○」は翻字・validation によって適切に処理できたものを,「翻字×」は処理できなかったものを指す。表 2 から，辞書に挙げられていない人名は，全般に単位数が多いことが分かる。

単位数	辞書あり	辞書なし	
		翻字○	翻字×
2	7	0	1
3	25	3	2
4	15	2	6
5	2	0	2
6	0	0	1
7	0	0	0
8	0	1	0
平均	3.2	4.2	4.0

表 2: 英語人名の単位数

ヒット件数	辞書あり	辞書なし	
		翻字○	翻字×
1,000 未満	7	3	7
1,000 以上 10,000 未満	6	1	1
10,000 以上 100,000 未満	18	1	3
100,000 以上	18	1	1
平均	232,746	43,686	31,891

表 3: 正解日本語訳の Google におけるヒット件数

定訳のヒット件数は表 3 のようになった。表 3 から，辞書に定訳が挙げられていない英語人名の方が，挙げられている英語人名よりもヒット件数が少ないことが分かる。ヒット件数が少ないということは，関係のない語がヒット件数の点で上回る可能性がそれだけ高いということである。即ち validation がうまく機能しない可能性が高いことを意味する。だが，先ほど見たように，辞書に定訳が挙げられていない英語人名は一般に単位数が多いので,「無関係だが語として存在する」文字列が生成される可能性は多少低いはずである。単位数ごとの平均ヒット件数を調べるなどして，今後 validation の可能性を検証したい。

6 おわりに

本研究では，翻訳者支援を目的として，英語人名に対して，辞書と翻字によって候補語を入手し，Web 検索によって日本語定訳を出力するシステムを提案した。今後は前章で挙げた改善方向を検証すると共に，NACSIS-CAT 中の著者人名や，Web ページから一定のパターンで抽出できる日英人名対を加えるといった，辞書の拡張を中心に研究を進めたい。また生没年・職業などの人物情報の援用も検討したい。

参考文献

[1] AbdulJaleel, Nasreen and Larkey, Leah S. (2003) “Statistical Transliteration for English-Arabic Cross Language Information Retrieval”, *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, p.139–146.

[2] Kang, In-Ho and Kim, GilChang (2000) English-to-Korean Transliteration Using Multiple Unbounded Overlapping Phoneme Chunks *Proceedings of the 17th Conference on Computational Linguistics*, p.418–424.

[3] Knight, Kevin and Graehl, Jonathan (1998) “Machine Transliteration”, *Computational Linguistics*, vol.24, no.4, p.599–612.

[4] Melamed, I. Dan (2000) “Models of Translational Equivalence among Words,” *Computational Linguistics*, vol. 26, no.2, p.221-249.

[5] EDP (2002)『100 万語収録のスーパー英和・和英辞典：英辞郎』アルク.

[6] 日外アソシエーツ (2002)『カタカナから引く外国人名綴り方字典』日外アソシエーツ.