対訳人名検索における翻字・サーチエンジンの有効性評価

辻慶太 (国立情報学研究所 人間・社会情報研究系：keita@nii.ac.jp)
佐藤理史 (京都大学情報学研究科：sato@pine.kuee.kyoto-u.ac.jp)
影浦峡 (国立情報学研究所 人間・社会情報研究系：kyo@nii.ac.jp)

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，翻字と Web に基づいて出力する手法を提案する。本手法は，英語人名をいくつかの日本語人名候補に翻字し，それらをサーチエンジンで検索してヒット件数の高い候補から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に利用することで，英語人名に関する適切な日本語訳あるいは「定訳」が出力できると考えている。本研究では，対訳辞書に収載されていない英語人名に焦点を当て，手法の有効性を検証した。

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を目的として，既存の対訳辞書にない英語人名に対し，適切な日本語訳語を得る手法を提案する。具体的には，まず英語人名をいくつかの日本語人名候補に翻字する。次にそれらをサーチエンジンで検索して，ヒット件数の高い候補を訳語として出力する（以下ではこれを「バリデーション」と呼ぶ）という手法を提案したい。翻字については，自然言語処理の様々な分野や言語対でその有効性が示されている（Collier et al. (1997)，Knight & Graehl (1997)，Kawtrakul et al. (1998)，Wan & Verspoor (1998)，Jeong et al. (1999)，Kang & Kim (2000)，Tsuji (2002)，AbdulJaleel & Larkey (2003)）。またサーチエンジンにおけるヒット件数を，表記の一般性を示す尺度として利用することで，人名に関する適切な訳あるいは「定訳」が出力できると考えている。以下では本手法の翻字とバリデーションについて述べ，次に実験の結果を提示する。

2 翻字とバリデーション

2.1 翻字

まず対訳辞書から翻字パターンを学習する方法について述べ，次に英語人名を翻字する手法について述べる。

辞書からの翻字パターンの学習は，辞書が挙げるカタカナ・英語の対に対して，日本語と英語を単位に分割することから始まる。まず日本語に関してはカタカナ１文字を１単位とした。ただし英語との対応を考え,「アイウエオァィゥェォャュョッー」については，直前のカタカナを子音と母音に分け，後者とつなげた形で１単位とした。例えば「マイ」は「ma，i」ではなく「m，ai」といった形の２単位に分割した（２つの母音をつなげたものを以下では長母音と呼ぶ）。また特に「クス」は“ x ”との対応を考えて１単位とした。以上，例えば「ニッコーリ」は「n，iTU，k，o-，ri」といった５単位に分割し,「フェリックス」は「h，uE，r，iTU，kusu」といった５単位に分割した。それに対して英語の方は，子音字と母音字に分け，それぞれを１単位とした（ただし同じ子音字が連続する場合はまとめて１単位とした）。その上で対応する日英単位列を DP マッチングで特定した。DP マッチングでは,「日本語側の k と英語側の c，k，ck が一致した場合は５点」といったヒューリスティックによるスコア表を用意しておき，各日英語対において最もスコアが高い対応付けを行った。そしてスコアが一定値以上の日英語対を，翻字パターンの学習用データとした。[1] 例えば “Niccoli”と「ニッコーリ」という日英語対は，単位同士の対応が高いので学習用データとして採用したのだが，この対からは「n=n，iTU=i，k=cc，o-=o，ri=li」といった翻字パターンを収集した。

翻字パターンの収集は以上の通りである。次に入力された英語人名に対する翻字の方法だが，これに関しては英語をまず先ほど同様の単位に分割した上で，単位のバイグラムに関する翻字パターン

[1] 学習に用いた後述の辞書には，表記が対応していないカタカナ・英語の対，例えば “oak”と「カシ」などが含まれる。スコアに閾値を設けるのは，このような対を学習用データから除外する為である。

として最も適切なものを適用した。例えば “Mike” は「m，i，k，e，mi，ke」といった単位に分割するが,「m，i」＝「マイ」というバイグラムの翻字パターンが,「mi，ke」＝「ミケ」といったパターンよりも学習データ中に多く存在することから，前者を適用するといった翻字を行った。

2.2 バリデーション

本研究の提案手法は，翻字によって生成された日本語訳候補をサーチエンジンで検索し，ヒット件数が高いものを日本語訳語として出力するものである。検索では，元の英語と日本語訳候補の AND 検索と，日本語訳候補のみの単独検索の２通りを行うことにした。例えば前者では，英語人名 “Douhet” に対し,「ドゥーエ」「ダウヘット」という日本語訳候補が翻字によって得られたならば，それぞれ「Douhet (AND) ドゥーエ」「Douhet (AND) ダウヘット」の形で検索を行い，ヒット件数の多い順に出力する。これは元の英語と共に現れている Web ページ数が多いほど，その日本語訳候補は訳語である可能性が高いという仮定に基づいている。サーチエンジンには Google を用いた。

3 辞書と調査対象人名

3.1 対訳辞書

対訳辞書としては，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』と EDP の『100 万語収録のスーパー英和・和英辞典：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，これらを本研究における学習用データとした。

3.2 調査対象人名

第１章で述べたような翻訳者支援に有効であることを目指し，本研究では政治・社会・コンピュータに関する Web 上の英語記事と，その日本語訳記事とから，英語人名とその日本語訳を収集した。具体的には，政治・社会分野の記事は *Counter Punch* や *Common Dreams*，*New York Daily News* などを用いた。これらはいずれも益岡・いけだらのブログ（http://humphrey.blogtribe.org/）や「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。コンピュータ分野に関しては，internetnews.com の記事を用いた。それらは IT media（http://www.itmedia.co.jp/）で日本語訳が提供されている。各英語人名に対する正解日本語訳だが，本研究では上記日本語ページに現れている訳や，辞書が挙げる訳などを参考に，筆者らが総合的に判断して正解を決定した。結果，401 個の英語人名が得られ，１つ当たりの正解日本語訳数は 2.2 であった。これら 401 個の英語人名は，前節の辞書に載っている 280 語と載っていない 121 語に分けて，調査対象とした。第１章で述べたように，本研究では主に後者の適切な扱いが課題となる。

4 結果

翻字の性能に関しては，まず各英語人名に対して，最も翻字として確からしい日本語人名候補 30 個を出力し，順位 N（$1 \leq N \leq 30$）の候補までに１つ以上の正解人名が含まれている確率（割合）を調べた。結果は図 1 のようになった。図の X 軸は 1 位から 30 位を，Y 軸はその順位までに正解人名が含まれている確率を表している。“DICT-P” は辞書に載っている英語人名の結果を，“NO-DICT-P” は辞書に載っていない英語人名の結果を表している。

図 1 から，例えば辞書に載っていない英語人名に関しては，日本語人名候補の上位５個までに１つ以上の正解人名を含む割合は約 61 %であること，下の順位の候補まで見ても，正解人名を含む割合は 66 %までしか上がらないこと，などが分かる。学習用データになっているので，ある意味当然であるが，全体に辞書に載っている英語人名に対しては翻字の性能は高く，辞書に載っていない英語人名に対しては翻字性能は低いことが分かる。

翻字によって生成された日本語訳候補 30 個を Google でバリデーションした結果は図 2，3 のようになった。[2] 図 2 は辞書に載っている英語人名に関する結果，図 3 は載っていない人名に関する結果である。翻字の時と同様，ヒット件数順位 N（$1 \leq N \leq 30$）の候補までに１つ以上の正解人名が含まれる確率を示してある。グラフの “VALIDATION-A”，“VALIDATION-S” はそれぞれ 2.2 節で述べた AND 検索，単独検索による結果であり，“NONE” はバリデーションを行わなかった場合の結果，即ち翻字のみの結果である。図 2 から例えば，日本語訳候補 30 個のうち，翻字スコア上位３位までの日本語訳候補に正解人名が含まれる確率は 83 %であったのに対し，Google での単独検索のヒット件数順に出力した場合は，上位３位までの同確率は 84 %になること，また AND 検索のヒット件数順に出力した場合は、同確率は 89 %に上がることな

[2] ヒット件数が 0 になった日本語訳候補は，翻字スコアで順位付けを行った。

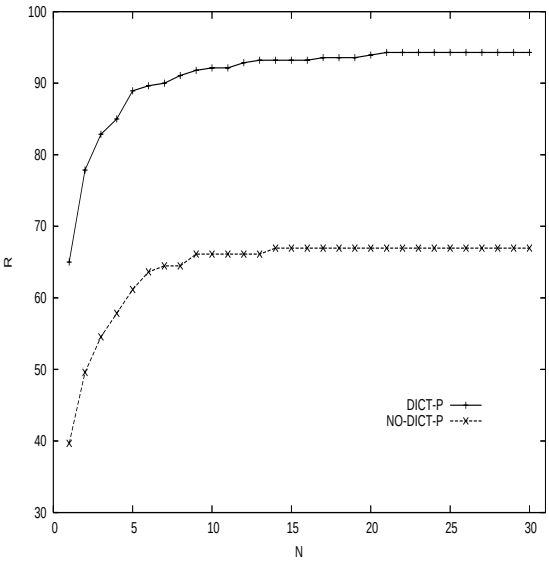

図 1: 翻字の結果

どが分かる。

辞書に載っている英語人名に関しては，AND 検索によるバリデーションは全般に有効であった。一方，辞書に載っていない英語人名に関しては，順位 1 位 2 位において有効であるものの，全体としてそれほど有効ではなかった。

5 誤り解析

5.1 翻字

翻字に関して観察された誤りとその対処策は以下の通りである：

非英語由来人名に対する誤り：“Sengupta”（セングプタ：インド系の名）の “gu”を英語流に「ガ」と翻字するなど，非英語由来の人名に対する翻字の誤りが観察された。これは学習データにこのような人名が相対的に少ないことに起因する。元となった言語ごとに学習データを分け，入力された英語人名に対して元の言語を推定し，その言語向けの翻字パターンを適用することが，今後の改善策として考えられる。その際，取り上げる言語としては，英語と文字体系が同じであるスペイン語・フランス語，文字体系が元々異なるアラビア語・ロシア語・中国語などを区別するのが効果的であろう。後者においては，例えば “Khudair”（クダイル：アラビア語由来の名）の “Khu”のように，特徴的な表記で，英語に無い音が表される場合が多く，それに向けた単位分割と翻字を行う必要がある。

語頭語末での誤り：今回は翻字において，語頭語末と語中とを区別しなかった。その為，例えば “Telesca”（テレスカ）“Tasca”（タスカ）における “-ca”を語中の一般的な翻字と同じように「キャ」として誤るケースが見られた。今後両者を区別する方向を検討したい。

促音・長音記号を余計に入れる誤り：翻字の結果には，“Buffett”（バフェット）を「バッフェット」としたり，“Jotiar”（ジョティアル）を「ジョティーアル」とするなど,「ッ」「ー」を余計に入れて誤っているケースが頻繁に観察された。「翻字スコアが 1 位の日本語訳候補から「ッ」「ー」を削除すると正解日本語訳が得られる英語人名の数」を調べたところ，辞書に載っていない英語人名においては 15 個，辞書に載っている英語人名においては 27 個あった。これらは翻字スコア 1 位の候補が誤っているケースのそれぞれ 20 %，27 %を占めた。2.1 節で述べたように，本研究の翻字パターンでは促音・長音記号には前の音の母音のみを組み合わせている。この為，学習データ中の例えば “ni”と「ニッ」，“ri”と「リッ」の両方から,「英語の “i”には日本語の「iTU」が対応すること」が学習される。その結果，入力英語中の “i”は，前の子音に拘わらず「iTU」に翻字される場合が多くなる。対処策としては，まず促音に関しては英語の後の方の子音と組み合わせることが考えられる。例えば “Rick”（リック）の “c”のように，促音に関与しているのは，後の英語子音であることが多い。

一方，長音記号の場合は，前後の英語子音が関与するケースは少ない為，促音の場合よりも対処が難しい。長音記号については,（2.1 節で述べた長母音に限らず）すべてのカタカナを子音と母音に分けて英語単位との対応を考えることで，問題が緩和される可能性がある。即ち，すべて子音と母音に分けることで，例えば英語の “i”に対応する日本語は “iTU”よりも “i”の方が多いことが適切に学習されるかもしれない。その場合，例えば “Abd”（アブド）のように英語に母音字 “aiueo”が現れていないケースに対応する為，英語においては NULL 文字を認め，それらと日本語母音との対応を考える必要がある。

5.2 バリデーション

辞書に載っている／いない英語人名共に,「単独検索によるバリデーションでヒット件数が 1 位になった語が正解日本語訳である割合」は,「翻字スコア 1 位の語が正解日本語訳である割合」よりも低くなっていた。これは例えば “Ahlam”に対して「アラーム」，“Catz”に対して「ケース」のように，翻字による日本語訳候補の中に，ヒット件数が非常に大きい一般語が偶然生じることによる（AND

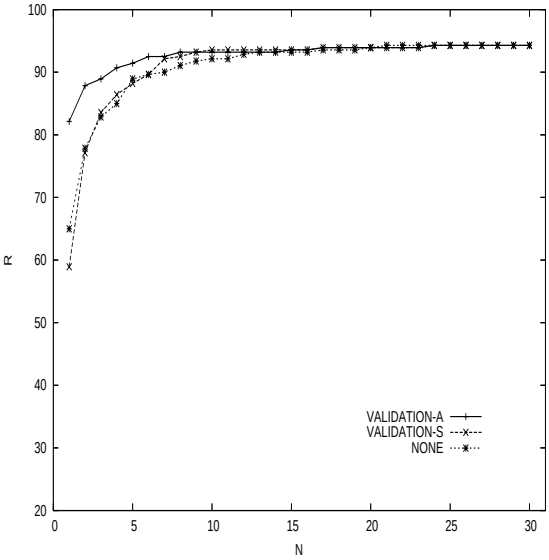

図 2: 辞書に載っている英語人名に関するバリデーションの結果

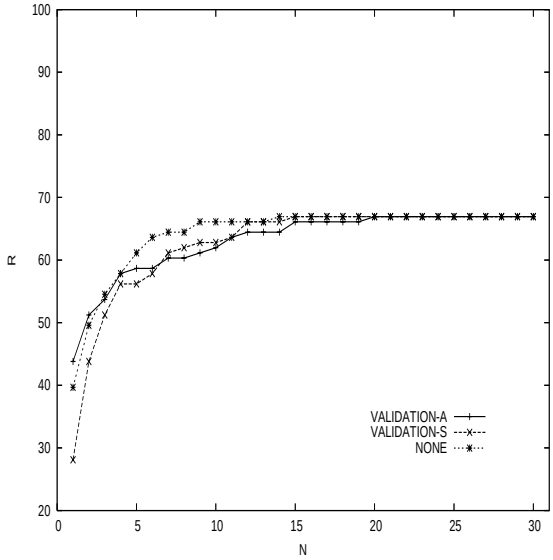

図 3: 辞書に載っていない英語人名に関するバリデーションの結果

検索の場合は，元の英語人名が組み合わさる為，このような一般語の順位アップは抑制される）。

さて，辞書に載っている英語人名 280 個のうち，正解日本語訳との AND 検索のヒット件数がすべて 0 であったものは 40 個（14.3 %）あった。また正解日本語訳の単独検索のヒット件数がすべて 0 のものはなかった。一方，辞書に載っていない英語人名 121 個においては，これらの値はそれぞれ 57 個（47.1 %），10 個（8.3 %）であった。即ち，辞書に載っていない英語人名の半数近くは，正解日本語訳との AND 検索ヒット件数がすべて 0 となる。従ってこれらに対しては，元々バリデーションはあまり機能しないことが予想される。

6 おわりに

本研究では翻訳者支援を目的として，英語人名に対する日本語訳語を，翻字と Web に基づいて出力する手法を提案した。調査の結果，辞書に載っている英語人名に比べ，載っていない英語人名を適切に扱うことの難しさが示された。今後は前章で挙げた改善方向を検証すると共に，他のリソースからの情報の援用，例えば Web ページから自動抽出できる人名の利用や，発音辞書の利用といった研究を進めたい。特に後者は，英語に固有の音節の区切り方や発音のパターンが収集できる為，有望な方向と考えている。さらには生没年や職業などの個人情報の用い，人物ごとの伝統的な訳し分け（“Hepburn”に対する「ヘプバーン」と「ヘボン」）などにも対処できるようにしたい。

参考文献

[1] AbdulJaleel, N. & Larkey, L. 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] Collier, N. et al. (1997) “Acquisition of English-Japanese proper nouns from noisy-parallel newswire articles using KATAKANA matching,” *Proceedings of the NLPRS*, p.309–314.

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

[4] Jeong, K. S. et al. (1999) “Automatic identification and back-transliteration of foreign words for information retrieval,” *Information Processing and Management*, 35(4), p.523–540.

[5] Kang, I. & Kim, G. (2000) “English-to-Korean transliteration using multiple unbounded overlapping phoneme chunks,” *Proceedings of the 17th COLING*, p.418–424.

[6] Kawtrakul, A. et al. (1998) “Backward transliteration for Thai document retrieval,” *Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and Systems*, p.563–566.

[7] Knight, K. & Graehl, J. (1998) “Machine transliteration,” *Computational Linguistics*, 24(4), p.599–612.

[8] Tsuji, K. (2002) “Automatic extraction of translational Japanese-KATAKANA and English word pairs from bilingual corpora,” *International Journal of Computer Processing of Oriental Languages*, 15(3), p.261–280.

[9] Wan, S. & Verspoor, C. M. (1998) “Automatic English-Chinese name transliteration for development of multilingual resources,” *Proceedings of the COLING-ACL’98*, p.1352–1356.

[10] 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.