原言語の起源に基づく潜在クラス翻字モデル

萩原 正人　関根 聡
楽天技術研究所
{masato.hagiwara, satoshi.b.sekine}@mail.rakuten.co.jp

1　はじめに

翻字とは，“バラクオバマ／Barak Obama” のように，表記体系の異なる言語間における音韻的な翻訳である．外来語を取り入れる際に用いられる主要な方法の一つであり，日本語の固有名詞の表記ゆれの主な原因となっている．

語と語の翻字関係を捉える際には，音韻に基づく書き換えモデル [5] や，綴りの対応関係を統計的に学習する手法 [3] などが用いられる．しかし，外来語には異なる起源を持つ語が混在するため，単一の翻字モデルでは関係を捉えきれないという問題が生じる．例えば，フランス語起源の “piaget／ピアジェ” と英語起源の “target／ターゲット” の “get” のように，原言語が異なる場合，音韻・綴りの対応も異なる場合がある．

この問題に対して，言語・性別等の原言語の起源を明示的にモデル化・分類し，翻字モデルを切り替えて用いるクラス翻字法が提案されている [6]．この手法では原言語の起源をタグ付けした翻字関係にある語のペア（翻字ペア）からなる訓練集合が必要である．しかし，一般語に比べて翻字の必要性が高い固有表現に対しては特に，このように起源がタグ付けされた訓練集合を入手することは難しい．また，外来語の問題があるため，言語起源タグが翻字モデルに必ずしも役に立つとは限らない．例えば，“spagetti” という語はイタリア語起源であるが，英語の辞書にも含まれているため，この語に対して英語の音韻モデルを適用しても “スパゲッティ” との翻字関係を正しく捉えられない可能性がある．

そこで本研究では，原言語の起源を，直接観察できない潜在クラスとしてモデル化し，翻字ペアに対して尤もらしい翻字モデルを適用する潜在クラス翻字モデルを提案する．潜在クラスおよび対応する翻字モデルは，翻字関係にあるペアの集合から EM アルゴリズムにより求められる．これによって，例えば “spagetti／スパゲティ” にはイタリア語に対応する潜在クラスが結びつき，翻字関係を正しく認識できると期待される．

評価実験では，欧米諸語が混在した人名および固有名の翻字リストを用いて，外国語に対応する日本語の語を推定する翻字のタスクの精度を評価した．その結果，潜在クラスを用いない従来の翻字モデルと比較して同等以上の精度を示した．

関連研究として，テキスト読み上げシステムにおいて，原言語の起源を取り入れることにより固有名詞の発音を改善できることが報告されている [7]．また，本手法に類似したものとして，クエリログから文字の誤り確率を EM アルゴリズムにより求める手法 [1] も提案されているが，文字単位の誤り確率のみを扱っているため，本手法のほうがより一般的であると言える．

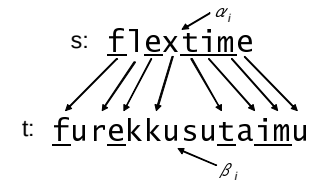

図 1: アルファベータ法における最小編集系列
下線はマッチ操作を表す

2　アルファベータ法

本研究における翻字の基礎モデルとして，翻字ペアの綴りの書き換え確率を直接モデル化するアルファベータ法 [3] を用いた．アルファベータ法は，文字の置換・挿入・削除それぞれの編集操作のコストを 1 とみなす通常の編集距離の一般化であり，$\alpha \to \beta(\alpha, \beta$ は長さ 0 以上の文字列) の形の文字列書き換え操作に対して確率値を与える．本モデルを用いて，語 s を語 t に書き換える確率は，以下のように求められる：

$$P_{AB}(t|s) = \max_{T \in \mathrm{Part}(t), S \in \mathrm{Part}(s)} \prod_{i=1}^{|S|} P(\alpha_i \to \beta_i), \quad (1)$$

ここで，$\mathrm{Part}(x)$ は語 x の全ての分割の集合であり，単一の分割 X とは，連結すると元の語になる部分文字列の系列 $X = (x_1, x_2, ...)$ のことである．上式は，語 s, t の全ての分割 $\mathrm{Part}(t)$, $\mathrm{Part}(s)$ から，書き換え確率を最大にする分割 T, S を見つける問題に相当する．全体の対数を取り，$-\log P(\alpha \to \beta)$ を文字列書き換え操作 $\alpha \to \beta$ のコストと見なすと，この問題は書き換えコストの合計の最小値を求める問題と等価である．よって，通常の編集距離の動的計画法と同様の計算により解くことができる．なお実際には，書き換え確率 $P(\alpha \to \beta)$ は，α の位置（語頭・語中・語末）によって条件付ける．すなわち，実際には $P(\alpha \to \beta|\mathrm{PSN}(\alpha))$, $\mathrm{PSN}(\alpha) \in \{$ 語頭, 中間, 末尾 $\}$ を扱うが，本稿では簡単のため略記する．

本モデルにおける書き換え確率 $P(\alpha \to \beta)$ は，翻字ペアからなる訓練集合から学習する．まず，通常の編

集距離を求める動的計画法を用いて，最短の編集操作の系列を求める．図 1 の場合，編集操作の系列は f → f, ε → u, l → r, e → e, ε → k, x → k, ... となる．次に，書き換えの文脈情報を含めるため，全ての非マッチ操作 ($\alpha \to \beta(\alpha \neq \beta)$ となるような操作) を，隣接する編集操作と併合し，書き換え対を作る．この時，書き換え対の文字列最大長を W に制限する．

例えば，$W = 2$ の場合，最初の非マッチ操作 ε → u は，左右 1 操作とそれぞれ併合され，f → fu, l → ur が作られる．次の非マッチ操作 l → r からは，l → ur, le → re が作られる．最後に，こうして併合されたものも全て含めた書き換え操作の出現頻度から，各書き換え操作の確率を求めることができる．なお，ある翻字ペアに対する最短編集系列は一意ではないため，全ての最短編集系列を考慮し，出現頻度の平均値を取る．

3 クラス翻字モデル

前節にて解説したアルファベータ法は，スペル訂正 [3], 翻字 [2], クエリ書き換え [4] 等のタスクにおいて優れた性能を示したという報告がある．しかし，このようにして訓練集合から学習された書き換えモデルは，訓練集合の統計情報を平均した単一的なものであり，1 節で示したような起源の異なる語に対して翻字モデルを切り替えて使うといったことができない．

Li. et al [6] は，この問題が中国語における逆翻字にも存在することを指摘している．“亜歴山大／ Alexsandra” など印欧諸語由来の名前の場合，中国標準語発音 (ピンイン) “Ya-Li-Shan-Da” に基づいた音韻モデルによって翻字を扱うことができるが，“山本／ Yamamoto” など日本語由来の場合，“Shan-Ben” という発音をそのまま対応させるのではなく,「漢字の日本語読み」というモデルを考慮する必要がある．

そこで彼らは，原言語の起源 l，性および姓／名の別 g，の 2 つの要因を考え，それらによって条件付けた翻字モデル $P(t|s,l,g)$ を線形結合し，$s \to t$ の翻字確率を以下のように求めるモデルを提案している．

$$P(t|s)_{\text{soft}} = \sum_{l,g} P(t,l,g|s) = \sum_{l,g} P(t|s,l,g)P(l,g|s) \quad (2)$$

ここで，言語の起源に関する要因 $c = (l, g)$ をクラスと呼ぶこととする．上式は翻字元 s に対するクラスの確率 $P(c|s)$ をまず求め，推定されたクラス c と翻字元 s に対する翻字先の確率 $P(t|s,c)$ の重み付き和により，翻字確率を求めていると解釈することができる．

なお，このモデルでは，全てのクラスを考慮しその重み付き和を求めている．これは，入力 s をクラスごとに確率値付きでソフトクラスタリングしていると見なすことができる．一方，入力 s に対して確率の最大となるクラス $c^* = \arg\max_{l,g} P(l,g|s)$ を用いて，

$$P(t|s)_{\text{hard}} \quad \propto \quad P(t|s,c^*) \quad (3)$$

として s から t への翻字確率を推定する手法（ハードクラスタリング）を用いることもできる．評価実験では，両者の手法を比較する．

Li et al. では，l(言語起源) は英語, 日本語，中国語ピンインの 3 つのいずれか，g(性別) は姓, 男性名, 女性名, の 3 つのいずれか，という設定を用いている．翻字の際の中国語漢字の出現・選択傾向は，言語起源，性別に強い影響を受けており，例えば姓については異なり文字の 19.2%が言語間で重複しているのみである．このクラス翻字モデルを考慮することによって，翻字の性能が向上することが報告されている．

4 潜在クラス翻字モデル

前節で述べたクラス翻字モデルでは，起源の異なる語に対して異なる翻字モデルを統合して用いているが，翻字ペアに対して明示的に起源がタグ付けされた訓練集合から，クラスの判定モデル c を構築する必要であるという制限がある．

そこで，各翻字ペアに対して明示的なクラス c を対応づけるのではなく，潜在的なクラスを表す確率変数 z を導入し，条件付き文字列書き換え確率 $P(\alpha \to \beta|z)$ を考える．この潜クラス z は，上記の言語起源や性別など，同じ翻字書き換え傾向を持つ翻字ペアのクラスに対応すると考えることができる．z は訓練集合からは直接観察されないが，以下のように EM アルゴリズムによって，訓練集合の尤度を最大化することにより繰り返し的に求めることができる．詳細な導出は紙面の都合上省略し，以下に EM アルゴリズムの更新式を示す．ここで，訓練集合 X_{train} は翻字ペアの集合 $\{(s_n, t_n)|1 \leq n \leq N\}$ であり，N は訓練集合に含まれる翻字ペアの数，K は潜在クラスの数である．

- パラメータ：

$$P(z = k) = \pi_k, P(\alpha \to \beta|z) \quad (4)$$

- E ステップ：

$$\gamma_{nk} \quad = \quad \frac{\pi_k P(t_n|s_n, z = k)}{\sum_{k=1}^{K} \pi_k P(t_n|s_n, z = k)}, \quad (5)$$

$$P(t_n|s_n, z) \quad = \quad \max_{T \in \text{Part}(t_n), S \in \text{Part}(s_n)} \prod_{i=1}^{|S|} P(\alpha_i \to \beta_i|z)$$

- M ステップ：

$$\pi_k^* \quad = \quad \frac{N_k}{N}, \quad N_k = \sum_{n=1}^{N} \gamma_{nk} \quad (6)$$

$$P(\alpha \to \beta|z = k)^* \quad = \quad \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \frac{f_n(\alpha \to \beta)}{\sum_{\alpha \to \beta} f_n(\alpha \to \beta)}$$

ここで，$f_n(\alpha \to \beta)$ は，2 節において述べた手法によって求められた，訓練集合の n 番目の翻字ペア中の書き換え対 $\alpha \to \beta$ の出現頻度である．以上により求められたクラス確率 π_k および 書き換え確率 $P(\alpha \to \beta|z)$ を用い，翻字確率を以下のように求める：

$$P_{\text{latent}}(t|s) = \sum_z P(t,z|s) = \sum_z P(z|s)P(t|s,z) \quad (7)$$

$$\propto \sum_z \pi_k P(s|z)P(t|s,z) \quad (8)$$

なお，潜在クラス翻字モデルのみでは，$P(s|z)$ を明示的にモデル化できないため，実際には $P(t|s,z)$ で近似した．ただし，この係数を完全に除外した場合でも，性能に及ぼす影響は 1.1%程度である．

5 実験

本節では，前節までに導入した各翻字モデルの性能評価実験について述べる．翻字モデルの性能は，与えられた翻字元 s' に対して，翻字モデル $P(t'|s')$ を用いて翻字先 t' をランキングする情報検索のモデルを用いて評価する．翻字ペアからなる評価集合 $X_{test} = \{(s'_n, t'_n)|1 \leq n \leq M\}$ を準備し，各 s'_n をそれぞれ検索要求，t'_n の集合を翻訳先の候補とした．翻字モデルを学習する際には，データセットを 5 等分した交差検定を実施した．評価指標としては平均逆順位 (Mean Reciprocal Rank; MRR) を用いた．MRR は，正解の順位の最小値を r とすると，$1/r$ を全検索要求に対して平均した値である．MRR を計算する際は，求められた翻字先候補上位 10 件のみを対象にした．

5.1 データセット 1：欧米人名リスト

本データセットは，欧羅巴人名録 [1] より抽出した欧米人名およびそのカタカナ読みから成り，ドイツ語 (de)，英語 (en)，フランス語 (fr) の人名それぞれ 2,470, 2,492, 1,747，合計 6,718 ペアが含まれる．なお，リストには姓・女性名・男性名の区別があるが，ここではそれらを全て区別せずに扱った．

英語以外の言語の人名にはアクセント記号が含まれるが，これらは言語判定のために有効な手がかりであるので，本データセットでは正規化せずにそのまま残した．大文字は小文字に統一した．

5.2 データセット 2：欧米固有名リスト

本データセットは，Wikipedia の言語間リンク (interwiki) を用いて抽出した，欧米固有名および対応する日本語表記からなり，ドイツ語 (de) 2,003, 英語 (en)5,530, スペイン語 (es) 781, フランス語 (fr) 1,918, イタリア語 (it) 1,091，合計 11,323 の固有名が含まれる．抽出には英語エントリと日本語エントリのタイトルの対応関係を用い，日本語エントリは，タイトルがカタカナ，長音記号，中黒のみから構成されるものだけを対象とした．ドイツ語，フランス語，イタリア語は，記事本文の第一文に “ドイツの”“フランスの”“イタリアの” がそれぞれ含まれているかどうかで判断した．ただし，スペイン語については，“スペインの”，“アルゼンチンの”，“メキシコの”，“ペルーの”，“チリの” のいずれかが，英語については，“アメリカの”，“イギリスの”，“オーストラリアの”，“カナダの” のいずれかが含まれていることを条件とした．

日本語・外国語ともに，タイトルの “,” と “(” 以降は除去した．また，日本語は中黒，外国語は “_” でチャンクに分割し，対応付けされたチャンク対を使用した．例えば，“Barak_Obama ／バラク・オバマ” というペアからは，“Barak ／バラク” と “Obama ／オバマ” の 2 つのチャンク対が生成される．チャンク数の異なるタイトル対は使用しなかった．また，外国語が長さが 2 以下のものから成るチャンク対は除外した．

本データセットでは，アクセント記号は全て正規化（ドイツ語の “ß” は “ss” に正規化）した．これは，データセット 1 に比べて難しい問題設定であるが，固有名を表記する際にアクセント記号が省略されることも多いため，より現実的な問題設定である．

[1] http://www.worldsys.org/europe/

表 1: 言語判定の結果 (データセット 1)

言語	de	en	fr
精度 (%)	80.4	77.1	74.7

表 2: 言語判定の結果 (データセット 2)

言語	de	en	es	fr	it
精度 (%)	65.4	83.3	48.2	57.7	66.1

5.3 実装の詳細

クラス翻字モデルの $P(c|s)$ としては，文字 3 グラムを用いた言語モデル確率を用いた．Witten-Bell discounting 法を用いてスムージングした．

日本語は，ヘボン式ローマ字に全て変換した．ただし，外来語を扱えるように，“wi ／ウィ” “we ／ウェ” などの若干の修正を加えてある．具体的な実装については，ソースコード [2] を参照のこと．長音記号 “—” については，直前の母音を重ねる．例えば，“スパゲッティー” は “supagettii” に変換される．

2 節で述べたアルファベータ法における書き換え対の文字列最大長は $W = 2$ とした．EM アルゴリズムの初期パラメータ $P(\alpha \to \beta|z)$ は，z の値にかかわらず，アルファベータ法により求められた確率値 $P(\alpha \to \beta)$ にガウスノイズを加えたものを用いた．また，π_k の初期値は $1/K$ に設定した．EM アルゴリズムの繰り返し回数は，予備実験の結果， 40 回に固定した．

5.4 結果

言語クラス判定 まず，クラス翻字モデルの $P(c|s)$ を用いて，式 (3) により言語判定した場合の精度を表 1, 表 2 に示す．言語判定の精度は，例えば Li et al. [6] と比較して総じて低いが，これは，印欧語系の名前の

[2] https://code.google.com/p/mhagiwara/source/browse/trunk/nltk/jpbook/romkan.py

表 3: 各モデルの性能比較 (%)

モデル	データセット 1	データセット 2
AB	94.8	90.9
HARD	90.3	89.8
SOFT	95.7	**92.4**
LATENT	**95.8**	**92.4**

場合，文字セットも同じであり，かつ，“Charles” が “チャールズ” もしくは “シャルル” となったり同一の名前もあるため難しいことを示唆している．データセット 2 では，クラスの数が増えているため，精度はさらに低下している．

なお，潜在クラス翻字モデルにおいても，翻字ペア (s_n, t_n) の γ_{nk} の値の最も大きくなる k^* を，対応するクラスと見なすことによって，訓練集合を教師なしクラスタリングすることができる．こうしてできたクラスタの純度 (purity) はおよそ 0.74 程度であった．

各翻字モデルの比較　次に，$P_{AB}(t|s)$ (AB), $P_{\mathrm{hard}}(t|s)$ (HARD), $P_{\mathrm{soft}}(t|s)$ (SOFT), $P_{\mathrm{latent}}(t|s)$ (LATENT) のそれぞれを用いて実際に翻字候補を獲得した結果を評価した．その結果を表 3 に示す．潜在クラス数は，元データの言語数と同じく，データセット 1 については $K = 3$，データセット 2 については $K = 5$ を用いている．LATENT は，SOFT と比較してほぼ同等の性能を示している．なお，以下に示すように，潜在クラス数の設定によっては LATENT の性能はさらに向上することがある．なお，クラス翻字法のクラス推定の精度が比較的低く，これが特に HARD の精度を下げる大きな原因になっていると思われる．一方，SOFT では，重みつき和によってクラス判定の誤りの悪影響が抑えられており，これが高い精度の原因であると考えられる．例えば，イタリア語とスペイン語の発音モデルは比較的似ており，$P(c = \mathrm{it}|s)$ と $P(c = \mathrm{es}|s)$ は両方とも類似した値になるため，結果的に両言語を混合したようなモデルが使われ，クラス判定の悪影響を受けないのだと思われる．

ここで，潜在クラスを用いない翻字法において，翻字先の推定に失敗した例を挙げる．データセット 1 においては，“Felix ／フェリックス” を “フィリス” , “Read ／リード” を “レアード” などと誤読する例が多く見受けられたが，これは，“x ／ックス”，“ea ／イー” という英語書き換えモデルが他の言語のモデルの影響を受けているためだと考えられる．LATENT では，これらの読みを正しく推定することができている．

データセット 2 において，潜在クラス翻字モデルを使い，改善されたペアには，“Caen ／カーン” (フランス語; “シャーン” と誤読) “Laemmle ／レムリ” (英語; “リアム” と誤読) “Xavier ／ザビア” (英語; “ガブリエル” と誤読) など，比較的変則的な発音の語に多い．最初の 2 語では，クラス翻字法におけるクラス判定が間違っているわけではないため，変則的な発音が，誤読の原因だと考えられる．一方で，“Hilda ／イルダ”(英語) を “ハルラ” と誤読するなど，さらに変則的な例では依然として翻字先推定に失敗する．

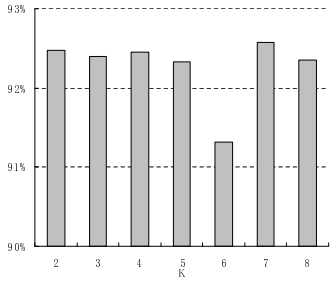

図 2: 潜在クラス数に対する翻字性能

潜在クラス数の影響　ここで，潜在クラスの数が翻字モデルの性能にどのような影響を与えるかを調べるために，k の値を 2 から 7 まで変化させ，データセット 2 の翻字性能がどのように変化するかを調べた．その結果を図 2 に示す．K の値の設定によっては，さらに性能が向上するが，$K = 6$ など大きな値になるとモデル $P(\alpha \to \beta|z)$ を推定するための統計が少なくなるため，EM アルゴリズムの初期値の影響が大きく不安定になっていると推測できる．

6　おわりに

本稿では，原言語の起源を潜在クラスとしてモデル化した潜在クラス翻字法を提案した．モデルの各パラメータは，複数の言語の起源の混在する訓練集合から EM アルゴリズムにより推定する．欧米の人名および固有名の翻字実験の結果，潜在クラスは言語の起源を表す明示的な情報を使っていないにも関わらず，従来手法と同等以上の性能を示した．

本モデルは，音韻情報などを用いず，文字列の置き換えのみにより翻字をモデル化しており，言語に依存しない汎用的なモデルである．今後，翻字元として，印欧諸語以外，および，翻字先として，日本語以外の言語を扱った翻字についても引き続き検討する．

参考文献

[1] Farooq Ahmad and Grzegorz Kondrak. Learning a spelling error model from search query logs. In *Proc. of EMNLP-2005*, pp. 955–962, 2005.

[2] Eric Brill, Gary Kacmarcik, and Chris Brockett. Automatically harvesting katakana-english term pairs from search engine query logs. In *Proc. NLPRS-2001*, pp. 393–399, 2001.

[3] Eric Brill and Robert C. Moore. An improved error model for noisy channel spelling. In *Proc. ACL-2000*, pp. 286–293, 2000.

[4] Masato Hagiwara and Hisami Suzuki. Japanese query alteration based on semantic similarity. In *Proc. of NAACL-2009*, p. 191, 2009.

[5] Kevin Knight and Graehl Jonathan. Machine transliteration. *Computational Linguistics*, Vol. 24, pp. 599–612, 1998.

[6] Haizhou Li, Khe Chai Sum, Jin-Shea Kuo, and Minghui Dong. Semantic transliteration of personal names. In *Proc. of ACL 2007*, pp. 120–127, 2007.

[7] Ariadna Font Llitjos and Alan W. Black. Knowledge of language origin improves pronunciation accuracy. In *Proc. of Eurospeech*, pp. 1919–1922, 2001.