# 意味訳型翻字システムにおけるユーザフィードバックの応用

黄 海湘† 藤井 敦‡
† 筑波大学大学院図書館情報メディア研究科
‡ 東京工業大学大学院情報理工学研究科

## 1 はじめに

科学技術や経済の発展に伴い，新しい専門用語や固有名詞が次々に作られている．また，これらの新語はインターネットによって世界中に発信される．そこで，外国の文化を取り入れるために，外国語の新語を迅速に母国語へ翻訳する必要性が高まっている．

外国語を翻訳する方法には「意味訳」と「翻字」がある.「意味訳」は原言語の意味を翻訳先の言語で表記する方法である.「翻字」は原言語の発音を翻訳先の言語における音韻体系で表記する方法である．固有名詞や専門用語は翻字されることが多い．

日本語や韓国語はカタカナやハングルなどの表音文字を用いて外国語を翻字する．それに対して，中国語には漢字しかないので，漢字を用いて翻字する．しかし，漢字は表意文字であるため，同じ発音に複数の文字が対応する．その結果，同音異義の問題が発生する．すなわち，翻字に使用する漢字によって，翻字された言葉に対する意味や印象が変わってしまう場合がある．

例えば，飲料水の名称である「コカコーラ（Coca-Cola）」に対して，様々な漢字列で発音を表記することができる．公式な表記は「可口可乐」であり，原言語と発音が近い．さらに,「可口」には「美味しい」,「可乐」には「楽しい」という意味があり，飲料水の名称として良い印象を与える.「Coca-Cola」の発音に近い漢字列として「口卡口拉」もある．しかし,「口拉」には「喉に詰まる」という意味があり，飲料水の名称としては不適切である．

また,「人名」や「地名」といった翻字対象の種別によっても使用される漢字の傾向が異なる．例えば,「宝」と「堡」の発音はどちらも/bao/である.「宝」には「宝物」や「貴重」などの意味があり中国語で人名や商品名によく使われるのに対して,「堡」には「砦」や「小さい城」などの意味があり中国語で地名によく使われる．

以上より，外国語を中国語に翻字する場合は，発音だけではなく，漢字が持つ意味や印象，さらには，翻字対象の種別（人名や企業名など）も考慮して漢字を選択する必要がある．本論文では，このような翻字を発音だけ考慮する翻字と区別して,「意味訳型翻字（semantic transliteration）」と称する．意味訳型翻字は，漢字を使う中国語や日本語などにおいて重要である．漢字圏への進出を計画する企業にとっては，企業名や商品名のネーミングにおいて意味訳型翻字の果たす役割が大きい．

翻字に関する既存の手法は,「狭義の翻字」と「逆翻字」に大別することができる．前者は外国語を移入して，新しい言葉を生成する [3, 5, 6, 7]．後者は既に翻字された言葉に対して原言語を特定する [1, 2]．逆翻字は主に言語横断検索や機械翻訳に応用されている．どちらの翻字も発音をモデル化して音訳を行う点は共通である．しかし，逆翻字は新しい言葉を生成しないため，本研究とは目的が異なる．本研究の目的は狭義の翻字である．以降，本論文では「翻字」を「狭義の翻字」の意味で使う．

中国語を対象とした翻字 [5] は固有表現の外来語に対して，発音モデルと言語モデルを組み合わせて使用する．しかし，翻字対象語の意味や印象を考慮していない．

Li ら [3] は外国人名を翻字する際に，対象人名の言語（日本語や韓国語など），性別，姓名を考慮した．しかし，この手法は人名のみを対象としているので企業名や商品名などには利用できない．Xu ら [6] は翻字対象語の発音と印象を考慮し，黄ら [7] は翻字対象の種別も考慮した．この 2 つの手法では，翻字対象の意味や印象を表す「印象キーワード」に基づいて，翻字に使用する漢字を選択する．しかし，生成した翻字の候補が多数存在するため，翻字結果の選択には時間と手間がかかる．

本研究は，黄ら [7] が構築した意味訳型翻字システムで出力する候補に対し，情報検索における適合性フィードバックを応用して，所望の翻字結果を効率良く探すシステムを提案する．

以下，2 節で提案するシステムについて説明し，3 節で本システムを評価する．

## 2 提案する翻字システム

### 2.1 概要

本研究で提案する翻字システムの概要は図 1 に示す．本システムでは，ユーザがウェブブラウザを通して翻字結果の閲覧およびフィードバックを行う．実際の翻字とフィードバック機能はサーバ側で行う．処理の流れは以下の通りである．

1. ユーザが翻字フォームのページに翻字対象語，翻字対象に関する印象キーワードを入力し，翻字対象の種別を選択する．なお，印象キーワードは Huang ら [4] の手法で自動的に収集することも可能である．

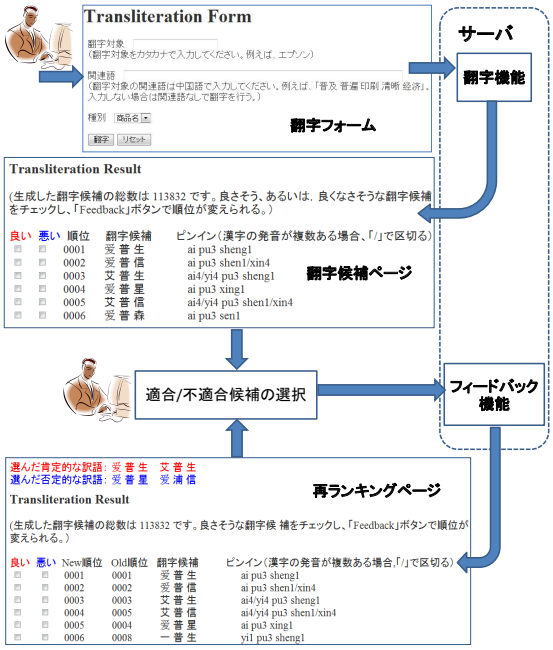


図 1: 提案する翻字システムの概要

2. 「翻字機能」が翻字候補リストをユーザに示す．
3. 示されたリスト中の各候補に対して，ユーザは「適合」あるいは「不適合」をチェックし，システムにフィードバックする．
4. 「フィードバック機能」が翻字候補を再ランキングし，その結果をユーザに示す．
5. ユーザが所望する翻字結果が見つかるまで上記 3. と 4. を繰り返す．

図 1 にある「翻字機能」は黄ら [7] の翻字手法を利用する．2.2 でその手法について概説する．「フィードバック機能」は本研究で新規に提案する部分であり，2.3 で説明する．

## 2.2 翻字機能

黄ら [7] で提案した翻字手法の概要を図 2 に示す．図 2 は，左から順番に「発音モデル」,「印象モデル」,「言語モデル」に大別される．以下，図 2 に基づいて説明する．

翻字機能への入力は 3 つある．1 つ目は，翻字対象となる外国語である．2 つ目は，翻字対象が指す実体や概念に対して，その印象を 1 つ以上の語（印象キーワード）で表現する．印象キーワードは中国語で入力する．3 つ目は，翻字対象の種別を「人名」,「企業名」などのカテゴリである．これらの入力に対して，1 つ以上の漢字列が翻字の候補として出力される．

図 2 の最左では,「発音モデル」によって，翻字対象が近い漢字列とそれぞれの確率が得られている．これらの漢字列が翻字候補となる．現在，翻字対象となる外国語は日本語のカタカナ語を対象としている．これはカタカナ語が発音表記であるローマ字に変換することが容易

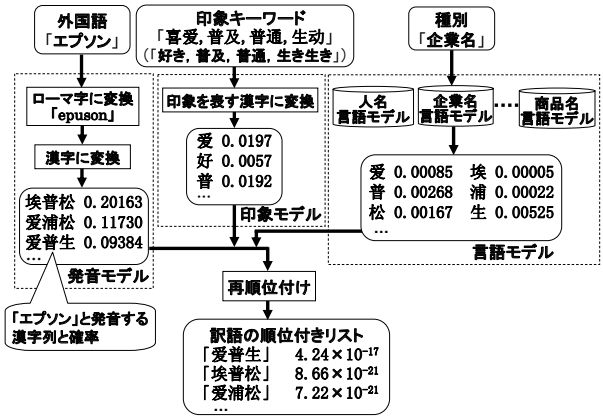


図 2: 図 1 における翻字機能の詳細

だからである．図 2 の中央では,「印象モデル」によって，印象キーワードに関連する漢字とそれぞれの確率が得られている．印象モデルを使用することによって，印象キーワードの中に含まれていない漢字も出力することが可能である．図 2 の最右では，種別に対応する言語モデルが選ばれる．

発音モデルで得られた翻字候補は複数になる場合があるため，それぞれに順位を付ける．具体的には，発音モデルで得られた翻字候補の確率による順位を，印象モデルと言語モデルで得られた漢字の確率によって再順位付けする．これは「翻字対象ローマ字表記 $R$」,「印象キーワード $W$」,「種別 $T$」が与えられた条件のもとで，$P(K|R,W,T)$ が最大になる漢字列 $K$ を選択することである．式（1）を用いて計算する．

$$P(K|R,W,T) \propto P(R|K) \times P(W|K) \times P(T,K) \quad (1)$$

式（1）の $P(R|K)$，$P(W|K)$，$P(T,K)$ はそれぞれ「発音モデル」,「印象モデル」,「言語モデル」と呼ぶ．以下，本研究で提案するフィードバック機能と最も関連している印象モデルと言語モデルの計算について説明する．

印象モデル $P(W|K)$ の計算では，$W$ と $K$ を単語 $w_i$ と漢字 1 文字 $k_j$ に分割して，$P(W|K)$ を $P(w_i|k_j)$ に基づいて近似する．

しかし，印象キーワードの数に制限はないので，$w_i$ と $k_j$ の数が常に同じであるとは限らない．そのために，式（2）を用いて $P(W|K)$ を計算する．各 $k_j$ については，$P(w_i|k_j)$ が最大となる $w_i$ だけを考慮する．

$$P(W|K) \approx \prod_j \max_i P(w_i|k_j) \quad (2)$$

なお，$P(w_i|k_j)$ は中国語漢字字典を用いて計算する．すなわち，漢字字典の見出し漢字を $k_j$ として，$k_j$ の意味記述で使用されている単語を $w_i$ とする．

言語モデル $P(T,K)$ は種別 $T$ に関するコーパスを用いて，式（3）で計算する．

$$P(T,K) = P(T) \times P(K|T) \propto P(K|T) \quad (3)$$

$P(T)$ は $K$ に依存しないので無視する．原理的には，種別 $T$ のコーパスが与えられた条件のもとで，漢字列 $K$ が生成される条件付き確率を計算する．

実際は，種別 $T$ に関するコーパスを用いて漢字の $N$ グラム確率を計算する．現在は，$N = 2$ として，式（4）で $P(K|T)$ を計算している．

$$P(K|T) \approx \prod_i P(k_i^{i+1}) \tag{4}$$

ここで，$k_i^{i+1}$ は $k_i$ と $k_{i+1}$ で構成される漢字バイグラムを表す．

### 2.3 フィードバック機能

黄ら [7] の翻字手法は翻字対象の意味や印象，および種別を考慮している．しかし，ユーザが所望する翻字結果が必ずしも翻字候補リストの上位にあると限らない．その際に，複数の翻字候補が出力されるので，翻字結果を効率的に選択する手段がない．最悪の場合は，すべての翻字候補を比較しなければならない．そこで，本研究は情報検索で用いられる適合性フィードバックの考えを取り入れ，翻字結果が候補リストの上位に上がるための「フィードバック機能」を提案する．

具体的には，2.2 で得られた翻字候補リストに対して，ユーザがいくつかの適合もしくは不適合の候補を選び，システムにフィードバックする．

ユーザのフィードバックは主に意味モデルと言語モデルに関係する．翻字ではどの候補も発音はそれほど変わらないためである．そこで，意味モデルと言語モデルに重みを掛け，全翻字候補に対して再ランキングする．具体的には，式（5）を利用して全翻字候補の確率を最計算する．

$$P(K|R,W,T) \propto P(R|K) \times P(W|K)^{\alpha} \times P(T,K)^{\beta} \tag{5}$$

式（2）と（4）を用いて，$P(W|K)^{\alpha}$ と $P(T,K)^{\beta}$ を展開すると，それぞれ式（6）と（7）になる．

$$P(W|K)^{\alpha} \approx \prod_j \max_i P(w_i|k_j)^{\alpha_j} \tag{6}$$

$$P(K|T)^{\beta} \approx \prod_i P(k_i^{i+1})^{\beta_i} \tag{7}$$

フィードバックされた「適合」候補の中に頻出する漢字の重みを大きくし，「不適合」候補の中に頻出する漢字の重みを小さくしたい．そこで，$\alpha_i$ と $\beta_i$ を式（8）によって計算する．

$$\alpha_i = \frac{1 + P_P(k_i)}{1 + P_N(k_i)} \qquad \beta_i = \frac{1 + P_P(k_i^{i+1})}{1 + P_N(k_i^{i+1})} \tag{8}$$

$P_P(\cdot)$ と $P_N(\cdot)$ は，それぞれユーザが「適合」もしくは「不適合」と判定した翻字候補における漢字の出現確率を表す．

なお，本フィードバック手法は「適合」か「不適合」の片方だけがフィードバックされた場合でも機能する．

## 3 評価実験

### 3.1 実験方法

本研究で提案した翻字システムの有効性を評価するために，ユーザフィードバックによる翻字結果の順位について調べた．

黄ら [7] の評価実験と同じデータを使用した．具体的には，翻字対象語として日中対訳辞書に登録されているカタカナ語から無作為に選んだ 210 語を使用した．翻字対象の印象キーワードは日本語が分かる中国人判定者 2 名に与えてもらった．また，判定者 2 名がそれぞれ翻字候補から選んだ訳語を正解訳語として使用した．

各翻字対象語に対して，判定者が判定した正解訳語から 1 つを選択し，「適合」の候補としてフィードバックする．判定者は正解訳語しか選んでいなかったため，評価実験では，「不適合」の候補に関するフィードバックは用いることができない．

また，「適合」の候補 1 つしかフィードバックしない場合，式（8）の重みの計算方法によって，フィードバックに用いた候補の順位は必ず元より向上，あるいは等しいであることが保証されている．そこで，判定者が 1 つ正解訳語しか判定していない用語は評価の対象外とする．

ユーザがフィードバックを行う際の手間を考慮して，フィードバックする正解訳語の順位が翻字候補リストの上位 10 以内とする．そこで，判定者が判定した正解訳語の順位が全て翻字候補リストの上位 10 以外の用語は評価の対象外とする．結果として，評価実験で使用した翻字対象語数を表 1 に示す．

表 1: 翻字対象語数の内訳

| 判定者 | 黄ら [7] の翻字対象語 | 実験対象語 | 正解訳語 |
|---|---|---|---|
| A | 210 | 61 | 187 |
| B | | 54 | 149 |

### 3.2 実験結果

表 2 にフィードバック機能を利用した結果を示す．表 2 の「フィードバックなし」と「フィードバックあり」は，フィードバックを使用しない場合とフィードバックを使用した場合それぞれにおける正解訳語の平均順位である．対象正解訳語の中に，フィードバックに用いた正解訳語自身は含まない．表 2 を見ると，判定者と関係なく，フィードバックを使用する正解訳語の平均順位は向上した．

図 3 は表 2 の結果に対して判定者 2 名の結果を合わせ，正解訳語の順位に関する分布を分析した結果である．図 3 を見ると，上位 1〜5 までに入った正解訳語数はフィードバックの有無によって変化がなかった．しかし，上位 6〜10 と 11〜20 までに入った正解訳語数ではフィードバックを使用する方が多かった．

以上をまとめると，中国語への翻字において，ユーザフィードバック機能が有効であった．

表 2: フィードバックの有無による正解訳語の平均順位

| 判定者 | フィードバックなし | フィードバックあり |
|---|---|---|
| A | 99 | 93 |
| B | 137 | 126 |
| 平均 | 118 | 110 |

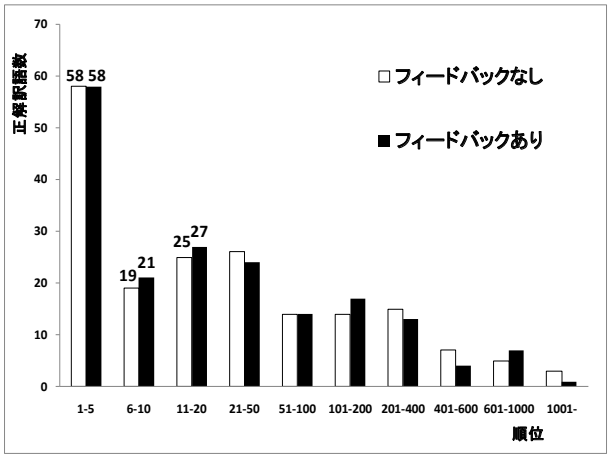


図 3: 正解訳語の順位分布図

表 2 の結果に対して，フィードバックの有無によって正解訳語の順位がどのように変化するかを考察した．結果を表 3 に示す．表 3 の「語数」は，図 1 にある「正解訳語」の中からフィードバックに用いた訳語を除いた語数である．表 3 の「向上」,「同等」,「低下」は，フィードバックによって順位が変化した件数の内訳である．

表 3: 正解訳語の順位変化

| 判定者 | 語数 | 正解訳語の順位変動 | | |
|---|---|---|---|---|
| | | 向上 | 同等 | 低下 |
| A | 107 | 45 | 46 | 16 |
| B | 80 | 40 | 34 | 6 |
| 合計 | 187 | 85 | 80 | 22 |

判定者 2 名の結果を合わせると，フィードバックなしに比べて，フィードバックありでは約 46%(85/187) の正解訳語は順位が向上し，約 11%(22/187) の正解訳語は順位が低下した．正解訳語の順位が低下した原因は主に以下の 2 通りであった．

(i) フィードバックに用いた訳語の順位が向上したために正解訳語の順位が低下した．この原因によって順位が低下した正解訳語は 7 件あった．例えば,「オメガ」では，フィードバックに用いた訳語「欧美佳」の順位が 8 位から 3 位に向上したため，正解訳語「欧美嘉」の順位が 6 位から 8 位に下がった．

(ii) フィードバックに用いた訳語と正解訳語に含まれた漢字があまり一致しなかった．この原因によって順位が低下した正解訳語は 15 件あった．例えば,「ナイキ」では，フィードバックに用いた訳語「耐克」と正解訳語「纳义基」に一致する漢字が 1 つもなかったため,「纳义基」の順位が 68 位から 70 位に下がった．

## 4 おわりに

中国語では，外国語を翻字するときに，表意文字である漢字を使用する．しかし，発音は同じでも使用する漢字によって翻字結果の意味や印象が異なる．本研究は，外国語を中国語に翻字するときに，既存の意味訳型翻字手法で得られた複数の翻字候補に対しユーザが所望する翻字結果を効率良く探すために，ユーザフィードバックを利用するシステムを提案した．また，評価実験で提案システムの有効性を示した．

今後の課題として，正解訳語の順位変動が小さかったため，フィードバック機能で利用する重みの計算については改良する必要がある．また,「不適合」の候補をフィードバックする実験も必要である．

## 参考文献


[1] Atsushi Fujii and Tetsuya Ishikawa, “Japanese/English cross-language information retrieval: Exploration of query translation and transliteration”. Computers and the Humanities, Vol.35, No.4, pp.389–420, 2001.

[2] Kevin Knight and Jonathan Graehl, “Machine Transliteration”. Computational Linguistics, Vol.24, No.4, pp.599–612, 1998.

[3] Haizhou Li, Khe Chai Sim, Jin-Shea Kuo, and Minghui Dong, “Semantic Transliteration of Personal Names”. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pp.120–127, 2007.

[4] HaiXiang Huang and Atsushi Fujii, “Effects of Related Term Extraction in Transliteration into Chinese”. In Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008), pp.643–648, 2008.

[5] Paola Virga and Sanjeev Khudanpur, “Transliteration of Proper Names in Cross-Lingual Information Retrieval”. In Proceedings of the ACL Workshop on Multilingual and Mixed-language Named Entity Recognition, pp.57–64, 2003.

[6] LiLi Xu, Atsushi Fujii, and Tetsuya Ishikawa, “Modeling Impression in Probabilistic Transliteration into Chinese”. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp.657–664, 2006.

[7] 黄 海湘, 藤井 敦, 石川 徹也, “中国語への翻字における確率的な漢字選択手法”. 電子情報通信学会論文誌, Vol.J90-D, No.10, pp.2914–2923, 2007.