中国語への翻字における関連語抽出の効果

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 〒305-8550 茨城県つくば市春日1-2
E-mail: {lectas21, fujii}@slis.tsukuba.ac.jp

あらまし
外国語の専門用語や固有名詞を翻字するときに、日本語や韓国語ではカタカナやハングルなどの表音文字を用いる。それに対して、中国語では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、発音は同じでも漢字によって意味や印象が異なる可能性がある。そこで、中国への翻字では適切な漢字を選択する必要がある。本研究は、翻字対象の関連語をWorld Wide Webから自動的に抽出し、翻字対象を表す印象キーワードとして利用する翻字手法を提案する。評価実験によって提案手法の有効性を示す。

Effects of Related Term Extraction in Transliteration into Chinese

HaiXiang Huang and Atsushi Fujii
Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2 Kasuga, Tsukuba-shi, Ibaraki, 305-8550, Japan
E-mail: {lectas21, fujii}@slis.tsukuba.ac.jp

Abstract
To transliterate foreign technical terms and proper nouns, in Japanese and Korean, phonograms, such as Katakana and Hangul, are used. In Chinese, the pronunciation of a source word is spelled out with Kanji characters. However, because Kanji comprises ideograms, different Kanji are associated with the same pronunciation, but can potentially convey different meanings and impressions. In this paper, we propose a method to select characters in transliteration into Chinese using related terms extracted from the World Wide Web. We show the effectiveness of our method experimentally.

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、インターネットによって世界中に発信される。外国の文化を取り入れるために、これら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。専門用語や固有名詞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し、文字によって意味や印象が異なる。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。

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。公式の表記は「可口可楽」であり、原言語と発音が近い。「可口」には「美味しい」、「可楽」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。「可口可楽」の発音に近い漢字列として「口可口拉」もある。しかし、「口可」には「喉に詰まる」という意味があり、飲料水の名称として不適切である。

別の例として、音楽家の「ショパン（Chopin）」は中国語で「肖邦」と表記する。「肖」は中国人名の苗字によく使われる漢字である。「肖」と同じ発音の漢字には「消」がある。しかし、「消」は「消す」や「消滅する」などの意味があるため、人名には不適切である。

以上の例より、外国語を中国語で翻字する場合は、発音だけでなく、漢字が持つ意味や印象、さらに、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ります。この点は、企業名や商品名を中国に普及させてブランドイメージを高める企業にとって特に重要である。

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[4, 5, 6, 7]。後者はすでに翻字された言葉に対して原言語を特定する
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している。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4, 5, 6, 7]は人名や地名などの外来語に対して、発音モデルと言語モデルを単独または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。

Xuら8]は翻字対象語の発音と印象を考慮した翻字手法を提案した。黄ら9]は翻字対象語の種別も考慮した翻字手法を提案した。これら2つの手法では、翻字対象の印象を表す「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、印象キーワードはユーザが中国語で与える必要がある。

本研究では、翻字対象語の関連語をWorld Wide Webから自動的に抽出し、印象キーワードとして使用する翻字手法を提案する。

以下、2.で本研究で提案する手法について説明し、3.で提案手法を評価する。

2. 提案する翻字手法
2.1 概要
本研究で提案する翻字手法の概要を図1に示す。図1は、左から順に、「発音モデル」、「印象モデル」、「言語モデル」に大別される。以下、図1に基づいて翻字手法について説明する。

図1 提案する翻字手法の概要
図1の最右では、入力された種別カテゴリに対応する言語モデルが選ばれる。図1では、「企業名モデル」が選ばれている。

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。

以下、2.2で印象キーワードの生成について説明する。2.3で確率的な漢字選択手法の全体像について説明する。2.4〜2.6で「発音」、「印象」、「言語」のモデル化について個別に説明する。なお、2.2が本研究で新規に提案する部分である。2.3〜2.6は筆者らが提案した翻字手法の全体を理解するために必要である。

2.2 印象キーワードの自動生成

翻字対象が指す実体や概念に対して、その印象を中国語で表記した語を「印象キーワード」と呼ぶ。

本研究では、翻字対象の関連語をWorld Wide Webから自動的に抽出し、印象キーワードとして利用する。

印象キーワードの自動生成は「関連語候補の抽出」と「印象キーワードの選択」の2段階に分けられる。図2に翻字対象が「エプソン」の場合、印象キーワードを自動生成するまでの概要を示す。図2の上部では、翻字対象に関連する関連語候補の抽出過程を示している。下部では、生成する印象キーワードの選択過程を示している。以下、それぞれについて、2.2.1と2.2.2で説明する。

2.2.1関連語候補の抽出

翻字対象の関連語を抽出するために、翻字対象に関する情報源が必要である。例えば、翻字対象が商品名であれば、その商品を紹介する文書や、翻字対象が企業名であれば、企業の理念やイメージに関する文書である。

このような情報源として、World Wide Web上のフリー百科事典「ウィキペディア（Wikipedia）」日本語版を利用した。2006年12月15日の時点では約30万の見出し語があり、一般名詞のほかに、人名、地名、企業名、商品名が登録されている。関連語候補の抽出は以下の手順に従って行う。

1. 翻字対象語をWikipediaで検索し、結果ページを取得する。図2の例では、翻字対象の「エプソン」をWikipediaで検索し、検索結果のページが得られる。

2. 取得した結果ページからHTMLタグを削除し、Gooで検索

3. 形態素解析の結果から、翻字対象に関連語候補として、名詞、形容詞を抽出する。ただし、「名詞-数」、「名詞-接尾-助数詞」、「名詞-副詞可能」、「名詞-非自立」、「名詞-代名詞」は削除した。

図2では、「好き」、「普及」、「普通」、「良い」などが関連語の候補として抽出されている。

2.2.2印象キーワードの選択

Wikipediaから抽出した名詞と形容詞の中には、翻字対象と関連がない語が含まれているため、翻字に使用する関連語を選択する必要がある。

本研究では、翻字対象と関連語候補間の相互情報量を計算して、関連語の選択を行う。相互情報量は式(1)を用いて計算する。

\[ I(X,Y) = \log \frac{P(X,Y)}{P(X) \times P(Y)} \] 

\( P(X) \)と\( P(Y) \)は単語\( X \)と\( Y \)の出現確率で、\( P(X,Y) \)は\( X \)と\( Y \)の同時確率である。ここでは、\( X \)は翻字対象であり、\( Y \)は関連語候補である。図2の例では、\( X \)は「エプソン」である。\( Y \)は「好き」、「普及」、「普通」、「良い」である。

図2 印象キーワード自動生成の概要

http://ja.wikipedia.org/wiki/
2.3 漢字選択ための確率モデル

本研究における漢字の目的は、「漢字対象ローマ字表記」の「印象キーワード」を選択することである。式(2)を用いて計算する。

\[
P(K | R, W, T) = \frac{P(R, W, T | K) \times P(K)}{P(R, W, T)}
\]

\[
= \frac{P(R | K) \times P(W | K) \times P(T | K) \times P(K)}{P(R, W, T)}
\]

\[
\approx P(R | K) \times P(W | K) \times P(T | K)
\]

式(2)の2行目はベイズの定理を用いた変形である。3行目で，ローマ字，ピンインの音節，漢字は独立であると仮定する。4行目で，P(R, W, T)はKに依存しない値である。最終的に，P(K | R, W, T)はP(R), P(W), P(T)のそれぞれとして計算される。それぞれ「文法モデル」，「言語モデル」と呼ぶ。

2.4 発音モデル

発音モデルは，中国語の漢字列Kが与えられた条件のもとで，ローマ字表記Rが生成される条件付き確率P(R | K)である。式(3)を用いて計算する。ローマ字表記はヘボン式を使用し，中国語のピンインYを中間言語として，中国語の漢字に変換する。

\[
P(R | K) = \prod_{i=1}^{n} P(r_i | y_i) \times \prod_{k=1}^{n} P(y_i | k_i)
\]

式(3)はそれぞれローマ字の音節，ピンインの音節，漢字1文字である。

例えば，漢字列「愛普生」が与えられた条件のもとで，ローマ字の音節「ai pu son」が生成される確率を計算する場合は，ピンインの音節「ai pu sheng」を中継してもらい，次のように計算する。

\[
P(\text{ai pu son} | \text{愛普生})
\]

= P(\text{ai pu son} | \text{ai pu sheng}) \times P(\text{ai pu sheng} | \text{愛普生})

なお，式(4)を用いて計算する。

\[
P(r_i | y_i) = \frac{F(r_i | y_i)}{\sum F(r_i | y_i)}
\]

\[
P(y_i | k_i) = \frac{F(y_i | k_i)}{\sum F(y_i | k_i)}
\]

式(3)はピンイン列の場合は式(4)を用いて計算する。

\[
P(r_i | y_i) = \frac{F(r_i | y_i)}{\sum F(r_i | y_i)}
\]

\[
P(r_i | y_i) = \frac{F(r_i | y_i)}{\sum F(r_i | y_i)}
\]

P(r_i | y_i) = \frac{F(r_i | y_i)}{\sum F(r_i | y_i)}

2.5 印象モデル

印像モデルは，漢字列Kが与えられた条件のもとで，印象キーワード列Wが生成される条件付き確率P(W | K)である。

WとKを単語wと漢字1文字kに分割して，P(W | K)をP(w | k)に基づいて近似する。

\[
P(W | K) = \prod_{j=1}^{n} \max P(w_j | k_j)
\]

式(3)は漢字3つと印象キーワード4つについてのP(w | k)を示している。表の「－」は，対応するwとkに対してP(w | k)が計算できないことを示す。
P(愛 喜 生 | 人) = P(愛 | 人) x P(喜 | 人) x P(生 | 人) = 0.02 x 0.03 x 0.03

\[ P(w_i | k_j) = \sum_{w_k} P(w_k | k_j) \]

(6)

\[ F(w_i, k_j) \]は愛を含む人名の見出し漢字を \( k_j \)として, \( k_j \)の意味記述に使用された単語 \( w_i \)とする.

実際, \( F(w_i, k_j) \)を計算するために, China語の漢字辞典から, 外来語の表記に良く使われる見出し漢字599文字を人手で選択し, 見出し漢字の意味記述をSuperMorphoで形態素解析して, 単語と見出し漢字の共起頻度を計算した. 表4は \( F(w_i, k_j) \)の例を示す.

| \( w_i \) | \( k_j \) | 普 | 普 | 生 | 生 |
|---|---|---|---|---|---|
| 喜愛 | 0.02 | 0.03 | 0.02 | 0.03 |
| 普及 | 0.01 | 0.03 | 0.02 | 0.03 |

表4: 漢字と単語との共起頻度に関する例

2.6 言語モデル

言語モデル \( P(T, K) \)は種別 \( T \)に関するコーパスを用いてモデル化する. 具体的には, 式(7)を用いて計算する.

\[ P(T, K) = P(T) x P(K | T) \approx P(K | T) \]

(7)

\( P(T) \)は \( K \)に依存しないので無視する. すなわち, 原理的には, 種別 \( T \)のコーパスが与えられた場合の \( K \)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, \( N=1 \)としている. 本研究では, 以下に示す3種類の言語モデルを構築し, 実験に使用した.

• 標準言語モデル: 中国北京大学「計算言語学研究所(Institute of Computational Linguistics)」が提供している「人民日报(人民日報) 1998年1月の新聞記事1ヶ月分から構築したモデルであり, 599文字を含む.」

• 名言語モデル: 上記「名言語モデル」が提供している「中文自然語言處理開放平台」から構築したモデルであり, 2,267件の漢字から構築した."

3. 評価実験

3.1 実験方法

本手法で提案した印象キーワード自動生成手法の有効性を評価するために, 人手で印象キーワードを与えた場合の翻訳精度と比較した. 具体的には, 以下のモデルの組合せ方について翻訳精度を比較した.

• 発音モデル, 言語モデル(音+言)
• 発音モデル, 印象モデル, 言語モデル: 印象キーワードを自動生成する(自動)
• 発音モデル, 印象モデル, 言語モデル: 印象キーワードを人手で与える(人手)

2番目の「自動」が本研究の提案手法である. 「音+言」と「人手」はそれぞれ期待される翻訳精度の下限と上限である. 本研究では, 3種類の言語モデルを構築した. しかし, 評価実験の目的は印象キーワード自動生成の有効性を評価することであるため, 使用する言語モデルは標準言語モデルに統一した. 実験に使う翻字対象として, まず, 日中対訳辞書[10]に登録されているカタカナ語1,140語から210語を無作為に選んだ. しかし,Wikipediaで検索しても結果が得られなかった場合と, 検索結果が複数の説明ページに対するリンクだけの場合は関連語抽出ができないため, 翻訳対象語から削除した. 最終的に210語のうち, 128語が残った. 翻字対象となっている128語に関する種別の内訳と例を表5に示す.

各翻字対象に対して, 日本語が分かる中国人被験者2名に印象キーワードを与えてもらった. 具体的には, 翻字対象の128語に対して, 日中対訳辞書[10]に記載された訳語を2名の中国人に示し, 意味を理解させた上で, 中国語で1語以上印象キーワードを与えてもらった. ただし, 被験者は Wikipediaを見ずに作業を行ったので, 被験者が与えた印象キーワードがすべてWikipediaに載っているとは限らない.

「人手」の翻字結果は, 各翻字対象について2名に対する正解訳語の順位を平均し, 上記全翻字対象を横断して順位を平均した."

翻訳精度を評価する尺度として, 「正解訳語の順位」を用いた. 日中対訳辞書[10]の訳語を「正解訳語」とした.
3.2 実験結果
式(2)に言語モデルの重みαを追加し、式(8)に変更し、αを変化させながら実験を行った。

\[ P(K | R, W, T) \propto P(R | K) \times P(W | K) \times P(T, K) \]  (8)

翻字の実験結果を表6に示す。表6において、1行目の「自動」は、自動的に生成した印象キーワードを利用した場合の結果である。2行目の「人手」は、人手で印象キーワードを与えた場合の結果である。

表6の「正解訳語の平均順位」は、3.1に示した②通りの組合せに対する結果をそれぞれに示している。ただし、正解訳語の平均順位が言語モデルの重みαを調整して、最も良い結果が得られたα=0.4に対する結果だけを示している。

表7より、「音+言」と比べると、印象モデルを加えた「自動」と「人手」の方が正解訳語の平均順位が高かった。しかし、「人手」の翻字精度よりも、「自動」の方が低かった。

3.3 考察
本手法で各翻字対象に対して、自動生成した印象キーワードを翻字に使用したときの正解訳語の内訳である。表7より、上位10と50までに入った正解訳語の数を見ると、「自動」と「人手」の差が僅かであった。

3.4 まとめ
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、中国語を翻字するときに、翻字対象に関する関連語をWikipediaから自動的に抽出し、翻字対象を表す印象キーワードとして利用する翻字手法を提案した。また、評価実験では、人手で与えた印象キーワードを使用する場合の翻字精度を比較した。その結果として、正解訳語の平均順位では、人手の235位に対して、自動抽出の場合は265位であった。さらに、上位10位までに入った正解訳語数を見ると、ほとんど差がなかった。この結果によって、本手法の有効性が示された。

しかし、Wikipediaの項目になった翻字対象には本手法を適用することができない。今回、これらの翻字対象を実験から削除した。実際に応用するとき、翻字対象の情報が得られる代替の情報源を探す必要がある。また、今後の課題としては、印象モデルの重み調整や関連語抽出のさらなる精緻化がある。

### 表5: 翻字対象に関する種別の内訳と例

| 種別   | 語数 | カタカナ語の例（中国語訳） |
|--------|------|--------------------------|
| 商品名 | 27   | アウディ（奥迪）          |
| 企業名 | 35   | サントリー（三得利）     |
| 地名   | 29   | スーダン（苏丹）         |
| 人名   | 13   | エンヤ（恩雅）           |
| 一般名詞| 24   | コーヒー（咖啡）         |

### 表6: 正解訳語の平均順位

|  | 音+言 | 自動 | 人手 |
|---|-------|------|------|
| 1~10| 43    | 36   | 22   |
| 11~50| 36    | 30   | 29   |
| 51~100| 12    | 10   | 9    |
| 101~500| 22   | 12   | 3    |
| 501~1000| 3    | 3    | 1    |
| 1000~| 2    | 1    | 1    |

### 表7: 順位階級における正解訳語数の内訳

|  | 順位階級  | 音+言 | 自動 | 人手 |
|---|-----------|------|------|------|
| 1~10| 43        | 48   | 49   |
| 11~50| 36        | 30   | 32   |
| 51~100| 12       | 10   | 8    |
| 101~500| 22      | 9    | 3    |
| 501~1000| 3       | 5    | 4    |
| 1000~| 2        | 1    | 3    |
### 表8: 自動抽出した印象キーワードが有効だった翻字対象の例

| 種別 | 翻字対象 | 正解語 | 手法 | 印象キーワード | 正解語順位 |
|------|-----------|--------|-----|----------------|------------|
| 商品名 | ボルボ | 自動 | 企業名 | カナ焼 | 812 |
| 人手 | 汽车 (デッド) 制 (デッド) 制 (デッド) 製 (デッド) 製 (デッド) 製 (デッド) 公司 (会社) 食品 (食品) 本多 (本多) 化学 (化学) 粉饰 (化粧) 买 (買収) 186 |
| 人手 | 汽车 (デッド) 制 (デッド) 制 (デッド) 製 (デッド) 製 (デッド) 製 (デッド) 公司 (会社) 食品 (食品) 本多 (本多) 化学 (化学) 粉饰 (化粧) 买 (買収) 186 |

### 表9: 自動抽出した印象キーワードが有効でなかった翻字対象の例

| 種別 | 翻字対象 | 正解語 | 手法 | 印象キーワード | 正解語順位 |
|------|-----------|--------|-----|----------------|------------|
| 人手 | 人手 | 一般 | 名词 | ミサ | 61 |
| 人手 | 人手 | 一般 | 名词 | ミサ | 61 |

### 手法

- 自動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
- 手動
参考文献

[1] Hsin-Hsi Chen, Sheng-Jie Huang, Yung-Wei Ding, and Shih-Chung Tsai. "Proper Name Translation in Cross-Language Information Retrieval". In Proceedings of the 36th Annual Meeting of the Association for Computational Conference on Computational Linguistics, pp.232-236, 1998.

[2] 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.

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

[4] ChunJen Lee and Jason S. Chang. "Acquisition of English-Chinese Transliterated Word Pairs from Parallel-Aligned Texts using a Statistical Machine Transliteration Model". HLT-NAACL 2003 Workshop: Building and Using Parallel Texts Driven Machine Translation and Beyond, pp.96-103, 2003.

[5] HaiZhou Li, Min Zhang, and Jian Su. “A Joint Source-Channel Model for Machine Transliteration”. Proceedings of ACL 2004, pp.160-167, 2004.

[6] 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.

[7] Stephen Wan and Cornelia Maria Verspoor. “Automatic English-Chinese name transliteration for development of multilingual resources". In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics, pp.1352-1356, 1998.

[8] LiLi Xu, Atsushi Fujii, and Tetsuya Ishikawa. “Modeling Impression in Probabilistic Transliteration into Chinese”. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, July 2006.

[9] 黄海湘, 藤井敦, 石川徹也. 中国語への翻字における漢字選択の手法. 電子情報通信学会技術研究報告, NLC2006, pp.7-12, Jul. 2006.

[10] 鈴木義明, 王文. 「日本語から引ける中国語の外来語辞典」, 東京堂出版, 2002.

[11] 新華字典電子版 v1.0.