2007－NL－177 (2)
2007／1／26

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「可口可乐」であり，原言語と発音が近い．「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，「口卡」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，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．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たい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

[1, 2, 3]. 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 1 に基づいて翻字手法について説明する．

本手法の入力は 2 つある．1 つ目は，翻字対象となる外国語である．2 つ目は，翻字対象の種別を「人名」，「企業名」，「商品名」などのカテゴリで入力する．そして，本手法はこれらの入力に対して，1 つ以上の漢字列を翻字の候補として出力する．

図 1 の最左では，「発音モデル」によって，翻字対象と発音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だからである．図 1 では，「エプソン（Epson）」を翻字対象の例として挙げている．ただし，ローマ字表記に変換できれば，他の言語を入力することも可能である．

図 1 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ない漢字も出力することが可能である．図 1 では，印象キーワード（「喜爱」，「普及」，「普通」，「生动」）に関連する漢字（「爱」，「普」，「好」など）とそれぞれの確率が得られている．

Xu ら[8]と黄ら[9]の研究では，印象キーワードはユーザが直接中国語で与えてもらう必要がある．したがって，ユーザは翻字対象が指す実体や概念について知らなければならない．また，中国語についても知らなければならない．これらの条件によって，システムを利用できるユーザが制限されてしまう．本研究では印象キーワードを自動生成して，この問題を解消する．

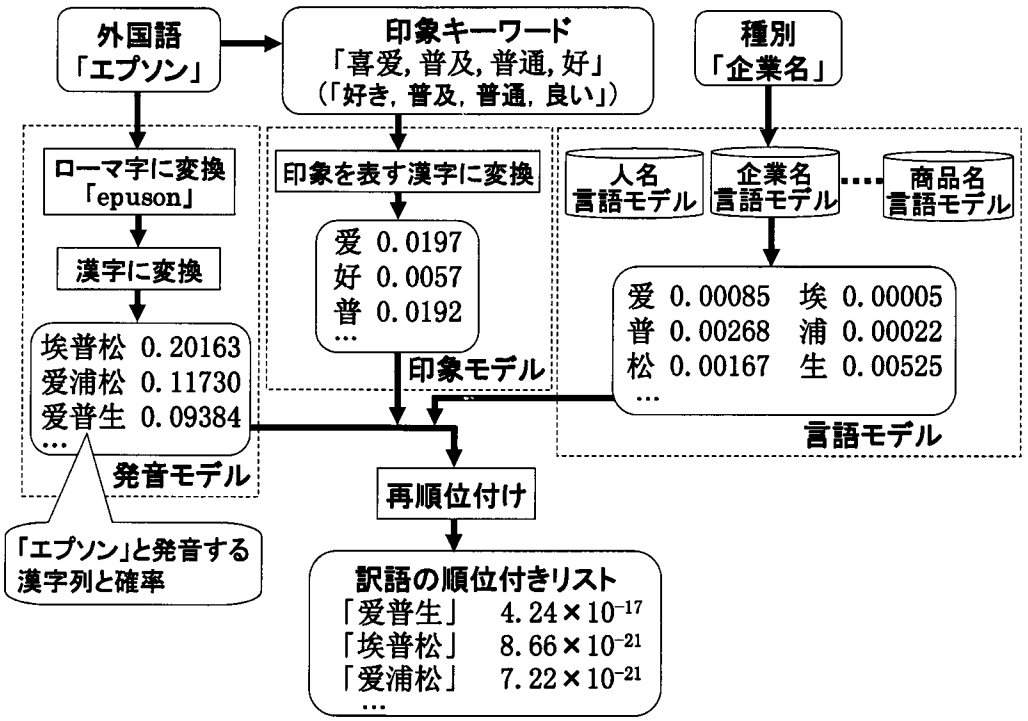

図 1 提案する翻字手法の概要

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．

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であり，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[1]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，茶筌で形態素解析を行う．
3. 形態素解析の結果から，翻字対象の関連語候補として，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」，「普遍」，「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．

$$I(X,Y)=\log\frac{P(X,Y)}{P(X)\times P(Y)} \qquad (1)$$

$P(X)$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」，「普遍」，「良い」である．

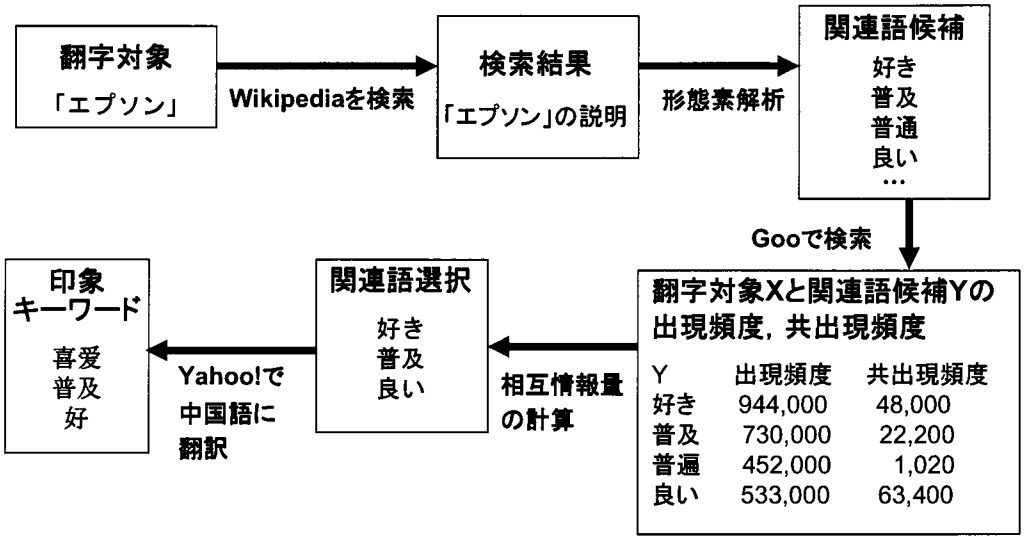

図 2 印象キーワード自動生成の概要

[1] http://ja.wikipedia.org/wiki/

印象キーワードの選択は以下の手順に従って行う．

1. $P(X)$，$P(Y)$，$P(X,Y)$を計算するために，“X”，“Y”，“X and Y”を検索キーワードとしてGoo[2]で検索し，検索結果の総数で近似する．
2. 翻字対象との相互情報量が上位 10 位までの候補を関連語として選択する．図 2 では，「エプソン」の関連語として「良い」，「好き」，「普及」が選ばれている．
3. 印象モデルに渡す印象キーワードは中国語であるため，翻字対象の関連語を中国語に翻訳する必要がある．今回は人手でYahoo! JAPAN[3]を利用して翻訳した．ただし，翻訳不能な語は削除する．図 2 では，「好」，「喜爱」，「普及」はそれぞれ「良い」，「好き」，「普及」に対する訳語であり，翻字対象の印象キーワードとして使用される．

2.3 漢字選択ための確率モデル

本研究における翻字の目的は，「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(2)を用いて $P(K/R,W,T)$を計算する．

$$
\begin{aligned}
&P(K \mid R,W,T) \\
&= \frac{P(R,W,T \mid K)\times P(K)}{P(R,W,T)} \\
&\approx \frac{P(R \mid K)\times P(W \mid K)\times P(T \mid K)\times P(K)}{P(R,W,T)} \\
&\propto P(R \mid K)\times P(W \mid K)\times P(T \mid K)\times P(K) \\
&\propto P(R \mid K)\times P(W \mid K)\times P(T,K) \qquad (2)
\end{aligned}
$$

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，R，W，Tが互いに独立であると仮定する．4 行目で，$P(R,W,T)$は K に依存しないため無視する．最終的に，$P(K/R,W,T)$は $P(R/K)$，$P(W/K)$，$P(T|K)$の積として近似される．それぞれ「発音モデル」，「印象モデル」，「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．

$$
\begin{aligned}
P(R \mid K) &\approx P(R \mid Y)\times P(Y \mid K) \\
&\approx \prod_{i=1}^{N} P(r_i \mid y_i)\times \prod_{i=1}^{N} P(y_i \mid k_i) \qquad (3)
\end{aligned}
$$

r_i，y_i，k_i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「爱普生」が与えられた条件のもとで，ローマ字の音節「e 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(e pu son | 爱普生)
$=P$(e pu son | ai pu sheng)$\times P$(ai pu sheng | 爱普生)
$=P$(e | ai)$\times P$(pu | pu) $\times P$(son | sheng) $\times P$(ai | 爱)
$\times P$(pu|普)$\times P$(sheng | 生)

式(3)中の$P(r_i/y_i)$と$P(y_i/k_i)$は式(4)を用いて計算する．

$$
P(r_i \mid y_i) = \frac{F(r_i,y_i)}{\sum_r F(r,y_i)}, P(y_i \mid k_i) = \frac{F(y_i,k_i)}{\sum_r F(y,k_i)} \qquad (4)
$$

$F(r_i,y_i)$はローマ字の音節r_iとピンインの音節y_iが共起する頻度であり，$F(y_i,k_i)$はピンインの音節y_iと漢字k_iが共起する頻度である．

これらの共起頻度を計算するために，日中対訳辞書[10]中のピンイン付き中国語と対応するカタカナ語 1,140 対を参考して，ローマ字とピンインの音節，ピンインの音節と漢字を人手で対応付けた．これらの一部をそれぞれ表 1 と 2 に示す．表 1 と 2 において，中国語のピンインには，発音の四声に基づいて 1～4 の識別子が付けられている．

ローマ字表記Rの分割が複数ある場合は，すべての可能な分割を考慮して$F(r_i,y_i)$を計算する．例えば，ローマ字表記「epuson（エプソン）」は 2 つのピンイン列を一致し，次のように分割される．

- e pu son : ai pu sheng
- e pu so n : ai pu son an

表1: ローマ字音節とピンイン音節の共起頻度に関する例

r_i	y_i	$F(r_i,y_i)$	r_i	y_i	$F(r_i,y_i)$
bi	bei4	2	ta	chou2	1
bi	bi3	17	ta	da2	8
bi	bi4	2	ta	da3	1

表2: ピンイン音節と漢字の共起頻度に関する例

y_i	k_i	$F(y_i,k_i)$	y_i	k_i	$F(y_i,k_i)$
ai4	爱	18	pu3	浦	3
ai4	艾	4	sheng1	生	3
pu3	普	28	sheng1	声	2

2.5 印象モデル

印象モデルは，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_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(5)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．

$$
P(W \mid K) \approx \prod_j \max_{w_i} P(w_i \mid k_j) \qquad (5)
$$

表 3 は漢字 3 つと印象キーワード 4 つについての$P(w_i|k_j)$を示している．表中の「—」は，対応するw_iとk_jに対して$P(w_i|k_j)$が計算できないことを示す．表 3 の例では，$P(W/K)$が次のように計算される．

[2] http://www.goo.ne.jp/
[3] http://honyaku.yahoo.co.jp/

P(喜爱 普及 普通 生动 | 爱普生)
=P(喜爱 | 爱)×P(普及 | 普)×P(生动 | 生)
=0.02×0.03×0.03

$P(w_i|k_j)$は式(6)を用いて計算する．

$$P(w_i \mid k_j) = \frac{F(w_i, k_j)}{\sum_w F(w, k_j)} \qquad (6)$$

$F(w_i, k_j)$はw_iとk_jの共起頻度である．本研究では，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の意味記述に使用された単語をw_iとする．

実際，$F(w_i, k_j)$を計算するために，中国語の漢字字典[11]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[4]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．

表3: $P(w_i|k_j)$の例

w_i	kj		
	爱	普	生
喜爱	0.02	——	——
普及	——	0.03	——
普通	——	0.02	——
生动	0.001	——	0.03

表4: 漢字と単語との共起頻度に関する例

kj	w_i	$F(wi, kj)$	kj	w_i	$F(wi, kj)$
普	普	13	生	生	91
普	普遍	6	生	生育	8
普	全面	1	生	盛	1

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．

$$P(T,K) = P(T) \times P(K \mid T) \approx P(K \mid T) \qquad (7)$$

$P(T)$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=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「计算语言学研究所(Institute of Computational Linguistics)」[5]が提供している「人民日报（人民日報）」1998 年 1 月の新聞記事一ヶ月分から構築したモデルであり，異なり漢字 4,540（延べ 12,229,563 漢字）を含む．
- 企業名言語モデル：中国科学院計算技術研究所が主催している「中文自然语言处理开放平台（中国語自然言語処理オープンソース）」[6]が提供している 22,569 社を含む「公司名录库（企業名リスト）」から構築したモデルであり，異なり漢字 2,267（延べ 78,432 漢字）を含む．
- 人名言語モデル：上記「中文自然语言处理开放平台」が提供している「带词性词频的扩展词典（品詞および出現頻度付き拡張辞典）」から 38,406 件の人名を抽出し，構築したモデルであり，異なり 2,318（延べ 104,443 漢字）を含む．

また，上記各モデルを構築する際に，SuperMorphoを用いて，コーパスの形態素解析を行い，句読点，記号，機能語を事前に削除した．

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字精度と比較した．具体的には，以下に示すモデルの組合せ方について翻字精度を比較した．

- 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字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，翻字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．

各翻字対象について，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]に記載された注釈を 2 名の中国人に示し，意味を理解させた上で，中国語で 1 つ以上の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全て 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，さらに全翻字対象を横断して順位を平均した．「自動」の翻字結果は，全翻字対象を横断して順位を平均した．翻字対象 1 つにつき，「自動」で生成した印象キーワード数は平均 8.9 であり，「人手」の場合は平均 6.6 であった．

翻字精度を評価する尺度として，「正解訳語の平均順位」を用いた．日中対訳辞書[10]の訳語を「正解訳語」とした．

[4] http://www.omronsoft.com/
[5] http://icl.pku.edu.cn/
[6] http://www.nlp.org.cn/

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）

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．

$$P(K \mid R,W,T) \propto P(R \mid K) \times P(W \mid K) \times P(T,K)^{\alpha} \quad (8)$$

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 6 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，「自動」の方が低かった．

表 7 は，各順位階級における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語数を見ると，「自動」と「人手」の差が僅かであった．

以上をまとめると，関連語抽出により自動生成した印象キーワードを用いることで中国語の翻字精度が向上した．また，人手で印象キーワードを与えた場合と同等の精度が得られた．

表6: 正解訳語の平均順位

音＋言	自動	人手
284	265	235

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順位を表 8 と 9 に例示する．表 8 は，人手で与えた印象キーワードとその翻字結果と比べ，自動生成した印象キーワードが有効だった翻字対象の例を示している．表 9 は，自動生成した印象キーワードが有効でなかった翻字対象の例を示している．

表 8 と 9 において，1 列目の「種別」は，翻字対象の種別である． 5 列目の「印象キーワード」は翻字対象に対する印象キーワードを示している．その中国語の意味を表す日本語を括弧内に示している．「自動」の場合はその中国語に翻訳する前の日本語を括弧内に示している．「人手」の場合は筆者が日本語訳を与えた．

6 列目の「正解訳語順位」では，「音＋言」，「自動」，「人手」を比較している．

表 8 と 9 の「印象キーワード」欄を見ると，自動的に抽出した印象キーワードは，人手で与えた印象キーワードとあまり一致していない．

表 9 を見ると，「シャネル」の場合は，自動抽出した印象キーワードが人手で与えた印象キーワードと一致する語はいくつかあった．しかし，より翻字対象の印象を表すような語がなかった．そのために，印象キーワードとして有効に機能しなかった．

「インテル」の場合は，印象モデルを加える前の「音＋言」の 11 位から 326 位まで低下した．自動生成した印象キーワードを見ると，「インテル」に関する印象を表す語がなかった．

「カタール」の場合は，自動生成した印象キーワードが全部カタール周辺国の国名となっている．「カタール」の印象キーワードとして適切ではなかった．

「モナリザ」と「ディスコ」の場合は，自動生成した印象キーワードの中に，翻字対象と関係のない語がいくつかがあった．例えば，「モナリザ」の「自己（自分）」，「手（手）」，「没有（無く）」であり，「ディスコ」の「好（良い）」，「回来（帰り）」，「制（製）」である．また，抽出した印象キーワード「非常」は，本来「とても」の意味なので，中国語の「大」ではなく，「非常」と訳すべきである．

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，外国語を中国語に翻字するときに，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と比較した．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 の項目になかった翻字対象には本手法を適用することができない．今回は，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7: 順位階級における正解訳語数の内訳

順位階級		1～10	11～50	51～100	101～500	501～1000	1001～3000	3001～10000	10000～
正解訳語数	音+言	43	36	12	22	9	3	2	1
	自動	48	30	10	29	3	5	1	1
	人手	49	32	8	30	4	3	1	1

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	正解訳語	手法	印象キーワード	正解訳語順位	
					音+言	音+印+言
商品名	ボルボ	沃尔沃	自動	标致汽车（プジョー） 雷诺（ルノー） 马自达（マツダ） 轿车（セダン） 小轿车（クーペ） 福特（フォード） 车型（車種） 福特汽车（フォード・モーター） 三菱汽车工业（三菱自動車） 全部的型号（フルモデル）	812	442
			人手	大型（大型） 舒适（心地よい） 身份（身分） 豪华（豪華） 快速（高速） 汽车公司（自動車会社） 安全（安全） 重型（重機） 卡车（トラック） 北欧（北欧） 瑞典（スウェーデン）		750
企業名	カネボウ	嘉娜宝	自動	花王（花王） 制（製） 制药（製薬） 化妆（化粧） 公司（会社） 食品（食品） 本多（本体） 肥皂（石鹸） 粉饰（粉飾） 收买（買収）	496	186
			人手	破产（破産） 化妆品（化粧品） 美容（美容） 美丽（綺麗） 娇小（きゃしゃで愛しい） 可爱（かわいい） 傲慢（傲慢） 娇丽（きゃしゃで綺麗） 气质（気質） 女孩（女の子） 才华（才気） 日本（日本） 产品丰富（製品が豊富） 食品（食品） 药物（薬）		236
地名	カラチ	卡拉奇	自動	美国（アメリカ） 国际（国際） 都市（都市） 郡（州） 机场（空港） 交通（交通） 滑冰场（リンク） 西面（西） 地域（地域） 隐私（プライバシー）	351	30
			人手	巴基斯坦（パキスタン） 魔幻（魔幻） 丰韵（あでやかな姿） 婀娜多姿（しなやかで美しい） 高贵（気高い） 清真寺（モスク） 伊斯兰教（イスラム教） 海港（港）		110
人名	カラヤン	卡拉扬	自動	柏林菲尔（ベルリン・フィル） 管弦乐（管弦楽） 交响乐团（フィルハーモニー） 伯恩斯坦（バーンスタイン） 交响曲（交響曲） 施特劳斯（シュトラウス） 马勒（マーラー） 瓦格纳（ワーグナー） 菲尔（フィル） 征尔（征爾）	17	10
			人手	指挥家（指揮家） 风范（風格） 奥地利（オーストリア） 音乐（音楽） 洒脱（さっぱりしている） 大方（気前がよい） 激情（激情） 豪迈（豪胆） 细腻（繊細） 精致（精緻）		23
一般名詞	ミサ	弥撒	自動	秘跡（秘跡） 圣体（聖体） 式司（司式） 典礼（典礼） 天主教（カトリック） 奉献（奉献） 主教（司教） 约翰（ヨハネ） 教会（教会） 语言（ことば）	61	19
			人手	神圣（神聖） 圣洁（純潔） 仪式（儀式） 端庄（荘重） 富丽（華麗） 天主教（カトリック教） 东正教（ギリシャ正教） 礼拜（礼拝） 信仰（信仰） 祭祀（祭祀） 宗教（宗教） 圣经（聖書） 周日（曜日） 庄严（厳粛） 圣餐（聖餐）神父（神父） 牧师（牧師）		23

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	正解訳語	手法	印象キーワード	正解訳語順位	
					音+言	音+印+言
商品名	シャネル	夏奈尔	自動	香水（香水） 附件（アクセサリー） 化妆（化粧） 便宜（安く） 名牌（ブランド） 固定（固定） 巴黎（パリ） 收集（コレクション） 黑色（ブラック） 衣服（服）	349	368
			人手	时尚（ファッション） 性感（セクシー） 野味（野生的） 魅力（魅力） 潮流（流行） 香水（香水） 服装（服） 奢侈品（贅沢品） 华贵（豪華） 大款（成金） 白领（ホワイトカラー） 名牌（ブランド） 靓丽（明るくて綺麗） 五号（５番） 法国（フランス） 帽子（帽子） 时髦（流行）		178
企業名	インテル	英特尔	自動	制（製） 各种各样（さまざま） 最合适地（最適） 大（非常） 装置（デバイス） 日语（日本語） 闪光（フラッシュ） 实现（実現） 法人（法人） 产品（製品）	11	326
			人手	爽朗（さわやか） 快乐（楽しみ） 男孩（男の子） 垄断（独占） 电脑（コンピュータ） 微软（マイクロソフト） 普及（普及） 网络（ネットワーク） 世界（世界） 用途广（用途が広い）		43
地名	カタール	卡塔尔	自動	科威特（クウェート） 也门（イエメン） 文莱（ブルネイ） 奥曼（オマーン） 几内亚比绍（ギニアビサウ） 利比亚（リビア） 约旦（ヨルダン） 巴林（バーレーン） 叙利亚（シリア） 黎巴嫩（レバノン）	317	173
			人手	阿拉伯（アラブ） 酋长（酋長） 石油（石油） 狭小（狭い） 沙漠（砂漠） 干燥（乾燥） 半岛电视台（アルジャジーラテレビ局） 战乱（戦乱） 混乱（混乱） 独立（独立） 有主见（自分の考えがある）伊斯兰（イスラム）		70
人名	モナリザ	蒙娜丽莎	自動	列奥纳多（レオナルド） 丽萨（リザ） 大（非常） 自己（自分） 微笑（微笑） 手（手） 美术馆（美術館） 没有（無く）	5535	4558
			人手	美丽（綺麗） 温柔（やさしい） 贵妇（貴婦人） 微笑（微笑み） 眼泪（涙） 神秘（神秘）少妇（若い奥さん）达芬奇（ダビンチ） 名画（名画）		3034
一般名詞	ディスコ	迪斯科	自動	好（良い） 音乐（音楽） 舞厅（ダンスホール） 大（非常） 朱莉安娜（ジュリアナ） 曲（曲） 回来（帰り） 制（製） 跳舞（ダンス）	93	60
			人手	消遣（気晴らし） 娱乐（娯楽） 跳舞（ダンス） 节奏（リズム） 快乐（楽しい） 热闹（賑やか） 五彩缤纷（色とりどりで美しい） 喧闹（騒がしい） 乱（乱れる） 啤酒（ビール） 美女（美女）		12

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 Data 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] 鈴木義昭，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.