社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－NL－174（9）
2006／7／28

中国語への翻字における漢字選択の手法

黄 海湘† 藤井 敦† 石川 徹也‡

†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
‡東京大学 史料編纂所・前近代日本史情報国際センター 〒113-0033 東京都文京区本郷 7-3-1
E-mail: †{lectas21, fujii}@slis.tsukuba.ac.jp, ‡ishikawa@hi.u-tokyo.ac.jp

あらまし 外国語の固有名詞や専門用語を翻字するときに，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．しかし，中国語では漢字を用いて翻字する．漢字は表意文字であるため，音は同じでも漢字によって与える印象が異なる．本研究は，中国への翻字において適切な漢字選択するために，発音だけでなく，翻字対象の印象や種別を考慮する手法を提案する．評価実験によって提案手法の有効性を示す．

キーワード 固有名詞，翻字，印象，言語モデル

Selecting Characters in Transliteration into Chinese

HaiXiang Huang† Atsushi Fujii† and Tetsuya Ishikawa‡

† Graduate School of Library, Information and Media Studies, University of Tsukuba 1-2 Kasuga, Tsukuba-shi, Ibaraki, 305-8550, Japan
‡ The Historiographical Institute, The University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan
E-mail: †{lectas21, fujii}@slis.tsukuba.ac.jp, ‡ishikawa@hi.u-tokyo.ac.jp

Abstract To transliterate foreign proper nouns and technical terms, in Japanese and Korean, phonograms, such as Katakana and Hangul, are used. However, in Chinese, Kanji characters are used to transliterate foreign words. Because Kanji characters are ideograms, characters that have the same pronunciation convey different meanings and impressions. We propose a method to select characters in transliteration into Chinese using pronunciations, impressions, and types of target words. We show the effectiveness of our method experimentally.

Keyword Proper nouns, Transliteration, Impression, Language models

1. はじめに

インターネットには新しい情報が様々な言語で発信されているため，外国語を迅速に翻訳する必要性が高まっている．

外国語の翻訳には「意味訳」と「翻字」がある。「意味訳」は原語の意味を翻訳先の言語で表現する方法である．「翻字」は原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，中国語は表意文字の漢字を用いて翻字する．表意文字は発音が同じでも文字によって意味や印象が異なるため，同音異義の問題が発生する．

例えば，飲料水「コカコーラ（Coca-Cola）」は中国語で「可口可楽」と表記する．原語と発音が近い上、「可口」は「美味しい」，「可楽」は「楽しい」という意味であり，飲み物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」は「喉に詰まる」という意味であり，飲み物の名称として不適切である．

また，音楽家の「ショパン（Chopin）」は中国語で「肖邦」と表記する．「肖」は中国人名で苗字として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」などの意味で使われるため，人名には不適切である．

以上の例から分かるように，外国語を中国語に翻字するときは、発音だけではなく，漢字が持つ意味や印象，翻字対象の種別（人名や企業名など）も考慮する必要がある．

自動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．前者は発音によって新しい語を生成する処理で，後者は既に翻字された言葉に対して元の外国語を特定する処理である．

本研究は狭義の翻字を対象とする．しかし，音訳をモデル化して言語間の変換を行う点で逆翻字の研究[1，2，3]とも関連がある．中国語を対象とした狭義

の翻字[4, 5, 6, 7]は音訳モデルと言語モデルを単独または組み合わせて利用する．しかし，漢字の意味や印象を考慮していない．また、翻字対象の種類別も考慮していない．

Xu ら[8]は発音と印象を考慮した翻字手法を提案した．本研究は Xu らの手法を拡張し，対象語の種別も考慮した翻字手法を提案する．

2. 本研究で提案する翻字手法

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 の破線で囲まれた①，②，③はそれぞれ「音訳モデル」，「印象モデル」，「言語モデル」である．入力する外国語はローマ字で表記できることが前提である．現在は，日本語のカタカナ語を対象としている．これは，カタカタ語をローマ字に変換することが容易だからである．ローマ字に変換できれば，他の言語も入力することが可能である．印象キーワードはその外国語に対する印象を中国語で表したものである．また，外国語の種別も入力する．

図 1 では，音訳モデルによって，「ビタミン」の発音に近い漢字列（「维塔命」，「维他命」，「韦他命」など）とそれぞれの確率が得られている．印象モデルによって，印象キーワード（「维护」，「他人」，「生存」）に関係する漢字（「维」，「卫」，「他」，「塔」，「命」など）とそれぞれの確率が得られている．言語モデルでは，種別（商品名）によってコーパスを選び，そのコーパスにおける漢字の出現確率が得られている．最後に，3 つのモデルで得られた確率を統合し，音訳モデルで得られた「维他命」や「维塔命」などの訳語候補が順位付けられて出力されている．

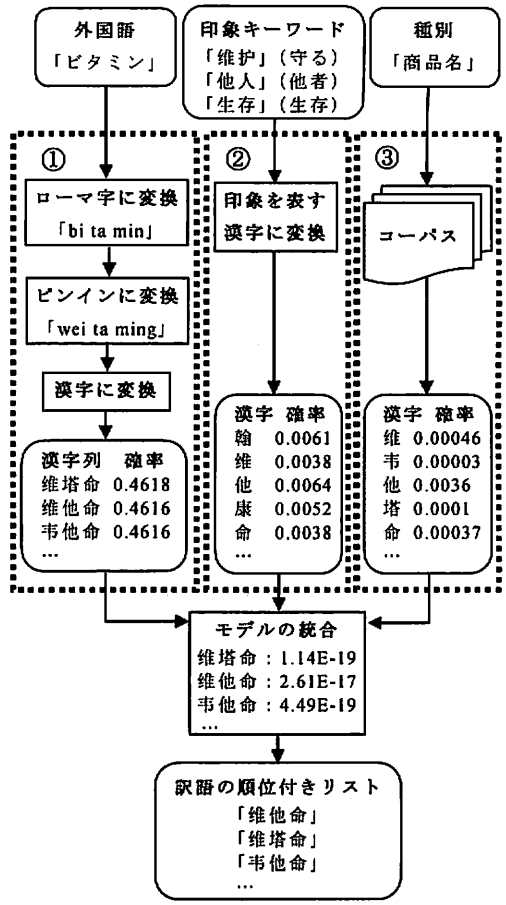

図 1：本研究で提案する翻字手法の概要

2.2 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，入力された言葉のローマ字表記 R と印象キーワード W が与えられたときに，漢字列 K を選択することである．これを式(1)に示す．

$$P(K \mid R,W) = \frac{P(R,W \mid K) \times P(K)}{P(R,W)} = \frac{P(R \mid K) \times P(W \mid K) \times P(K)}{P(R,W)} \propto P(R \mid K) \times P(W \mid K) \times P(K) \quad (1)$$

R, W は互いに独立であると仮定し，$P(R,W)$は K に依存しないため無視する．そして，$P(K|R,W)$に基づいて複数の漢字列候補に順位を付ける．式(1)の $P(R|K)$, $P(W|K)$, $P(K)$はそれぞれ「音訳モデル」，「印象モデル」，「言語モデル」に対応する確率である．Xu ら[8]は，「言語モデル」の $P(K)$を定数とし，漢字選択への影響を考慮していない．本研究はこの点を改良する．

2.2.1 音訳モデル

音訳モデル $P(R|K)$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で，式(2)により求める．ローマ字表記から漢字への変換は中国語のピンイン Y を中間言語として中継する．

$$P(R \mid K) = P(R \mid Y) \times P(Y \mid K) = \prod_{i=1}^{N} P(r_i \mid y_i) \times \prod_{i=1}^{N} P(y_i \mid k_i) \quad (2)$$

r_i, y_i, k_i はローマ字列，ピンイン列，漢字列をそれぞれ音節ごとに分割して得られたローマ字音節，ピンイン音節，漢字 1 文字である．

例えば，漢字列「维 他 命」が与えられた条件のもとでローマ字音節「bi ta min」が生成される確率を求める場合は，ピンイン音節「wei ta ming」を中継して，次のように計算する．

P(bitamin | 维他命)
=P(bi ta min|wei ta ming)×P(wei ta ming|维 他 命)
=P(bi|wei)×P(ta|ta)×P(min|ming)×P(wei|维)×P(ta|他)
×P(ming|命)

また，式(2)中の $P(r_i|y_i)$と $P(y_i|k_i)$は式(3)を用いて求める．

$$P(r_i \mid y_i)=\frac{F(r_i,y_i)}{\sum_r F(r,y_i)},\quad P(y_i \mid k_i)=\frac{F(y_i,k_i)}{\sum_y F(y,k_i)} \quad (3)$$

$F(r_i,y_i)$はローマ字 r_i とピンイン y_i が対応する頻度である．$F(y_i,k_i)$はピンイン y_i と漢字 k_i が対応する頻度である．これらの頻度を計算するために，日中対訳辞書[9]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，四声に基づいて1から4の識別子が付けられている．

表1: ローマ字音節とピンイン音節の対応頻度

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
bi	fo2	1	ta	da4	5
bi	pi2	1	ta	de2	2
bi	wei1	2	ta	tal	4

表2: ピンイン音節と漢字の対応頻度

y_i	k_i	$F(y_i,k_i)$
命	ming4	1
他	tal	6
韦	wei2	2
维	wei2	29
薇	wei2	1

表1の1列目は，ローマ字音節 r_i であり，2列目はピンイン音節 y_i である．3列目は r_i と y_i の対応頻度 $F(r_i,y_i)$である．表2の1列目は，ピンイン音節 y_i であり，2列目は漢字 k_i である．3列目は y_i と k_i の対応頻度 $F(y_i,k_i)$である．

2.2.2 印象モデル

印象モデルの $P(W|K)$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である．これを式(4)に示す．

$$P(W \mid K)=\prod_i P(w_i \mid k_i)=\prod_i \frac{F(k_i,w_i)}{\sum_w F(k_i,w)} \quad (4)$$

k_i は字典の見出し漢字であり，K を構成する漢字の1文字である．w_i は k_i の意味記述に現れる単語であり，W を構成する単語の1つである．$F(k_i,w_i)$は w_i が k_i の意味記述に使用された頻度である．

$F(k_i,w_i)$を計算するために，中国語の漢字字典[10]を使用した．字典から外来語の表記に良く使われる見出し漢字 599 件に対して，見出し漢字の意味記述をSuperMorpho（オムロンソフト社）で形態素解析し，抽出した単語と見出し漢字の共出現頻度を求めた．表3はその一部を示す．

表3: 漢字と意味記述で用いられた単語との共出現頻度

k_i	w_i	$F(k_i,w_i)$	k_i	w_i	$F(k_i,w_i)$	k_i	w_i	$F(k_i,w_i)$
命	命令	6	他	男女	1	维	绳	5
命	表示	2	他	男性	1	维	如	11
命	口	2	他	第三者	2	维	维	22
命	人	4	他	他人	1	维	保持	1
命	任命	3	他	他	19	维	纤维	1
命	命	38	他	的	5	维	维护	1

2.2.3 言語モデル

言語モデル $P(K)$はコーパスにおける漢字のユニグラム確率で計算する．ただし，対象語の種別ごとにコーパスを変更して，言語モデルを適応させる．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．

- 一般モデル：NTCIR-3 CLIR タスク中国語コレクションの新聞記事 2 年分をコーパスとして使用し，構築したモデルである．当コーパスは繁体字で書かれている．しかし，音訳モデルを構築するときに使った日中対訳辞書[9]と印象モデルを構築するときに使った漢字字典[10]は簡体字を使用しているため，コーパスをあらかじめ簡体字に変換した．句読点や符号などを除き，異なり 6,788 個の漢字（延べ 212,615,897 字）とコーパスにおける漢字の出現頻度のペアで言語モデルを構築した．
- 企業名モデル：中国科学院計算技術研究所数字化室＆軟件室が主催している「中文自然語言処理開放平台（中国語自然言語処理オープンソース）」が提供している約 22,569 社を含む企業名リストから構築したモデルである．具体的には，異なり 2,167 字（延べ 78,432 字）とリストにおける漢字の出現頻度で言語モデルを構築した．
- 人名モデル：上記「中文自然語言処理開放平台」が提供している「帯詞性詞頻的拡展詞典（品詞および出現頻度付き拡張辞典）」より 38,406 件の人名を抽出し，構築したモデルである．具体的には，異なり 2,318 字（延べ 104,443 字）と漢字の出現頻度で言語モデルを構築した．

「一般モデル」が対象語の種別を考慮していないのに対して，「企業名モデル」と「人名モデル」は対象語の種別を考慮している．

3. 評価実験

3.1 実験方法

本手法の有効性を評価するために，以下に示す 4 通りの手法を比較した．

- 音訳モデルのみ（R）
- 音訳モデルと印象モデル（$R+W$）
- 音訳モデルと言語モデル（$R+K$）
- 音訳モデルと印象モデルと言語モデル（$R+W+K$）

評価実験用の日本語は日中対訳辞書[9]に登録されているカタカナ語 1,140 語から 210 語（内訳：商品名 64 件，企業名 48 件，地名 36 件，一般名詞 41 件，人名 21 件）を選び，当辞書の訳語を正解とした．印象キーワードは日本語が分かる中国人 3 名に与えてもらった．

評価実験に使用したカタカタ語の例，正解訳語，判定者が与えた印象キーワードを表 4 に示す．

表4: 評価実験に使用したカタカナ語の例

カタカナ語	正解訳語	印象キーワード（括弧内は日本語訳）			種別
		判定者A	判定者B	判定者C	
アウディ	奥迪	轿车(乘用车) 富贵(富贵) 品质(品质) 速度(速度)	车名(车名) 德国(ドイツ) 豪华(豪華) 气派(気概) 价格(価額)	高贵(高貴) 速度(速度) 德国(ドイツ)	商品名
エプソン	爱普生	印刷机(印刷機) 知名(知名) 品质(品質) 优异(優秀)	电脑(パソコン) 打印机(プリンタ) 公司(会社) 产品(製品) 日本(日本)	喜爱(好き) 普及(普及) 生动(活き活き)	企業名
エンジェル	安琪儿	天使(天使) 可爱(可愛い) 幸福(幸せ) 爱心(優しい心)	神话(神話) 浪漫(ロマンチック) 天空(空) 白色(白色) 美丽(綺麗)	天使(天使) 平安(平安) 可爱(可愛い) 儿童(児童)	一般名詞
カネボウ	嘉娜宝	美丽(綺麗) 化妆品(化粧品) 皮肤(皮膚) 女人(女性)	化妆品(化粧品) 美容(美容) 皮肤(皮膚) 保护(保護) 营养(栄養)	女孩(女の子) 好(良い) 宝贝(宝)	企業名

表5: 印象モデルと言語モデルによる翻字結果の例

カタカナ語（正解訳語）	種別	判定者	R	R+W	R+K	R+W+K
サントリー（三得利）	企業名	A	7210	2109	86870	2103
		B		579		581
		C		71		72
サントメ（圣多美）	地名	A	694	105	77068	104
		B		110		107
		C		6		7
リスト（李斯特）	人名	A	733	1702	221	1662
		B		1042		1012
		C		210		201
モーター（马达）	一般名詞	A	133	92	52	89
		B		130		124
		C		40		38
コルゲート（高露洁）	商品名	A	4811	115	57311	115
		B		115		115
		C		105		105

3.2 印象モデルおよび言語モデルの有効性

対象語の種別を考慮せずに，一般モデルを使用して実験を行った．判定者 3 名それぞれの R，$R+W$，$R+K$，$R+W+K$ に関する翻字結果を抜粋して表 5 に示す．さらに，正解訳語の平均順位を表 6 に示す．1 つの対象語には正解が 1 つだけ存在し，210 語に対する正解の順位を平均した値を「平均順位」と呼ぶ．

表 6 では，$R+W$ は R より正解訳語の平均順位が向上し，最も良い結果であった．この結果より印象モデルの有効性が確認された．

表6: 正解訳語の平均順位

判定者	R	R+W	R+K	R+W+K
A	1732	265	22089	556
B		249		471
C		120		308
平均		211		445

表 7 は，$R+W$ と比べた場合に，$R+W+K$ に対する正解訳語の順位変動を示している．

表7: R+Wと比べた場合のR+W+Kにおける正解訳語の順位変動

判定者	向上	等価	低下
A	91	31	88
B	92	33	85
C	64	61	85

表 7 では，判定者 3 人ともに順位が低下した訳語の件数は向上した件数とほぼ同じであった．順位が低下する原因の一つは言語モデルを考慮する重みにあると考え，言語モデルの重みについて考察した．

言語モデルの重みを変化させるために，式(1)の対数を取り，式(5)のように言語モデルの項に重み α を追加した．

$$\log P(K \mid R,W) \propto \log P(R \mid K) + \log P(W \mid K) + \alpha \times \log P(K) \quad (5)$$

α の値を変化させながら実験を行った．結果を表 8 に示す．表 8 では，α の値として代表的な値だけを示し，「$R+W+\alpha K$」のように表記する．表 8 の「$R+W$」は表 6 と同じである．

表8: αの変化によるR+W+αKの結果

判定者	R+W	R+W+0.08K	R+W+0.09K	R+W+0.1K	R+W+0.2K
A	265	235	234	233	245
B	249	212	210	209	209
C	120	106	105	106	118
平均	211	184	183	182	191

表 8 では，$R+W$ より $R+W+\alpha K$ の方が良い結果となった．特に α が 0.1 の場合に，平均順位は約 182 位になり，$R+W$ の 211 位よりも良い結果だった．これより，翻字における言語モデルの有効性が確認できた．

今回の評価実験で使った 210 のカタカナ語は企業名，地名，人名，一般名詞と商品名の種別に分けられる．表 9 はその一部を示している．判定者それぞれの翻字結果を種別ごとに集計した結果は表 10 にまとめている．

表9: 一般モデルによる翻字結果の例

カタカナ語（正解訳語）	種別	判定者	R+W	R+W+0.1K	R+W+0.2K	R+W+0.3K
サントリー（三得利）	企業名	A	2109	2072	2114	2585
		B	579	777	976	1215
		C	71	100	252	474
サントメ（圣多美）	地名	A	105	112	136	215
		B	110	103	134	212
		C	6	16	27	55
リスト（李斯特）	人名	A	1702	1191	748	455
		B	1042	726	426	210
		C	210	76	31	20
モーター（马达）	一般名詞	A	92	85	77	68
		B	130	111	104	99
		C	40	35	33	29
コルゲート（高露洁）	商品名	A	115	98	83	64
		B	115	98	83	64
		C	105	88	72	57

表10: 一般モデルによる結果の種別による内訳

判定者	類別	R+W	R+W+0.1K	R+W+0.2K	R+W+0.3K
A	企業名	350	314	293	318
	地名	232	133	110	105
	人名	203	186	176	174
	一般名詞	89	89	127	282
	商品名	352	335	383	559
B	企業名	352	316	299	312
	地名	221	112	93	90
	人名	124	107	101	103
	一般名詞	87	81	109	216
	商品名	332	298	305	410
C	企業名	173	153	143	149
	地名	142	81	71	71
	人名	46	33	29	35
	一般名詞	91	101	139	241
	商品名	110	111	142	243

表 10 より，言語モデルの効果が翻字対象の種別ごとに異なることが分かる．例えば，判定者 B に関して，翻字対象が企業名の場合は，$R+W$ の 352 位より 299 位（$R+W+0.2K$）にしか向上していない．しかし，地名の場合は $R+W$ の 221 位から 90 位（$R+W+0.3K$）まで向上した．この結果より，翻字対象の種別を考慮して言語モデルを使い分ける必要がある．この点について，3.3 節でさらに実験を行う．

3.3 対象の種別における言語モデルの有効性

対象語の種別による言語モデル適応の有効性を考察するために，企業モデルと人名モデルを使用した．公平性を保つために，企業モデルと人名モデルを構築するときに，評価実験で使われた正解訳語をモデルから除いた．式(5)を用いて，一般モデルのときと同様に，α の値を変化させた．

3.3.1 企業名モデルの評価

企業名モデルを使用した翻字結果の一部を表 11 に示す．また，対象語の種別ごとに集計した結果を表 12 に示す．

表11: 企業名モデルによる翻字結果の例

カタカナ語（正解訳語）	種別	判定者	R+W	R+W+0.4K	R+W+0.5K	R+W+0.6K
サントリー（三得利）	企業名	A	2109	601	399	276
		B	579	115	85	64
		C	71	13	12	13
サントメ（圣多美）	地名	A	105	20	18	17
		B	110	19	17	15
		C	6	5	5	5
リスト（李斯特）	人名	A	1702	651	610	596
		B	1042	543	532	544
		C	210	86	82	79
モーター（马达）	一般名詞	A	92	51	43	41
		B	130	81	70	61
		C	40	13	13	11
コルゲート（高露洁）	商品名	A	115	85	98	112
		B	115	86	99	121
		C	105	59	65	75

表12: 企業名モデルによる結果の種別による内訳

判定者	種別	R+W	R+W+0.3K	R+W+0.4K	R+W+0.5K	R+W+0.6K
A	企業名	350	236	216	204	202
	地名	232	114	121	133	157
	人名	203	94	88	84	86
	一般名詞	89	59	65	76	97
	商品名	352	248	259	302	379
B	企業名	352	255	240	229	226
	地名	221	100	109	119	140
	人名	124	72	71	73	78
	一般名詞	87	72	83	99	124
	商品名	332	213	209	230	288
C	企業名	173	126	119	117	123
	地名	142	95	105	116	140
	人名	46	23	22	22	25
	一般名詞	91	67	71	82	98
	商品名	110	104	126	166	226
平均	企業名	291	205	192	183	183
	地名	198	103	111	123	146
	人名	125	63	60	60	63
	一般名詞	89	66	73	86	106
	商品名	265	189	198	233	298

表 10 より，一般モデルを使用した場合は，$\alpha=0.2$ のときに企業名の翻字において良い結果が得られ，正解訳語の平均順位はそれぞれ 293，299，143 であった．一方，表 12 より，企業名モデルを使用した場合は，判定者 A と B は $\alpha=0.6$，判定者 C は $\alpha=0.5$ の場合に良い結果が得られ，正解訳語の平均順位はそれぞれ 202，226，117 であった．

さらに，表 12 の「平均」を見ると，対象語の種別によらず，$R+W$ よりも $R+W+\alpha K$ の方が良い結果となった．しかし，企業名の向上が最も顕著であった．以上より，一般モデルよりも企業名モデルの方が企業名の翻字に効果的だった．

3.3.2 人名モデルの評価

人名モデルを使用した翻字結果の一部を表 13 に示す．また，対象語の種別で集計した結果を表 14 に示す．

表13: 人名モデルによる翻字結果の例

カタカナ語（正解訳語）	種別	判定者	R+W	R+W+0.4K	R+W+0.5K	R+W+0.6K
サントリー（三得利）	企業名	A	2109	1358	1238	1195
		B	579	375	372	364
		C	71	65	86	99
サントメ（圣多美）	地名	A	105	63	61	68
		B	110	61	60	67
		C	6	7	7	8
リスト（李斯特）	人名	A	1702	25	7	1
		B	1042	27	8	1
		C	210	2	1	1
モーター（马达）	一般名詞	A	92	36	27	24
		B	130	61	52	41
		C	40	12	7	6
コルゲート（高露洁）	商品名	A	115	106	116	136
		B	115	101	109	129
		C	105	82	95	105

表14: 人名モデルによる結果の種別による内訳

判定者	種別	R+W	R+W+0.3K	R+W+0.4K	R+W+0.5K	R+W+0.6K
A	企業名	350	327	339	361	410
	地名	232	72	76	95	152
	人名	203	55	45	40	38
	一般名詞	89	65	77	97	130
	商品名	352	205	230	287	390
B	企業名	352	341	354	374	410
	地名	221	56	60	72	119
	人名	124	28	22	20	20
	一般名詞	87	48	57	72	100
	商品名	332	168	172	204	284
C	企業名	173	208	227	255	297
	地名	142	49	57	83	150
	人名	46	13	11	11	11
	一般名詞	91	75	87	105	132
	商品名	110	73	97	143	216
平均	企業名	291	292	307	330	372
	地名	198	59	64	83	140
	人名	125	32	26	24	23
	一般名詞	89	63	73	92	121
	商品名	265	149	166	211	297

表 10 より，一般モデルを使用した場合は、α =0.2 のときに人名の翻字において良い結果が得られ，正解訳語の平均順位はそれぞれ 176，101，29 であった．一方，表 14 より，人名モデルを使用した場合は，α =0.6 のときに良い結果が得られ，正解訳語の平均順位はそれぞれ 38，20，11 であった．

さらに，表 14 の「平均」を見ると，企業名を除いて，対象語の種別によらず，$R+W$ よりも $R+W+\alpha K$ の方が良い結果となった．しかし，人名の向上が最も顕著であった．また，対象語が地名のときも向上が顕著であった．これは，人名に使われる漢字は地名でもよく使われるためである．以上より，一般モデルよりも人名モデルの方が人名の翻字に効果的だった．

4. まとめ

本研究は，外国語を中国語に翻字するときに，音訳モデル，印象モデル，種別ごとの言語モデルを使用した漢字選択の手法を提案した．また、評価実験によって印象モデルと種別ごとに適応させた言語モデルの有効性を示した．今後の課題は，印象キーワードを自動的に収集することである．

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] 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.

[10] 新華字典電子版　v1.0.