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

2006－NL－171（17）
2006／1／13

言葉の印象に基づく自動翻字手法

シュー　リーリー　　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　つくば市春日1－2
E-mail: {xulili,fujii,ishikawa}@slis,tsukuba.ac.jp

あらまし　他国の文化や技術をいち早く取り入れるために、外国語を母語に翻訳する必要がある。固有名詞や専門用語は翻字されることが多い。日本語や韓国語で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。中国語では漢字を用いて外国語を翻字する。しかし、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。本研究は発音と印象の両方を考慮して、外国語を中国語に翻字する手法を提案する。また、実験によって提案手法の有効性を示す。

Impression-based Automatic Transliteration

Lili XU,　Atsushi FUJII,　Tetsuya ISHIKAWA
Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
E-mail: {xulili,fujii,ishikawa}@slis.tsukuba.ac.jp

Abstract　To adopt foreign cultures and technologies rapidly, it is necessary to translate foreign words into a native language, for which proper nouns and technical terms are often transliterated. In Japanese and Korean, phonograms, such as Katakana and Hangul, are used to transliterate foreign words. In Chinese, Kanji characters are used to transliterate foreign words. However, because Kanji characters are ideograms, characters that have the same pronunciation convey different impressions. We propose a method to transliterate foreign words into Chinese using　pronunciations and impressions. We show the effectiveness of our method experimentally.

1. はじめに

インターネットの普及によって、世界中から多種多様な情報が入ってくるようになった。そこで、外国語を母語に翻訳して他国の文化や技術等を取り入れる必要性が益々高まっている。固有名詞や専門用語などは翻字されることが多い。日本語や韓国語における翻字は、カタカナやハングルなどの表音文字を用いて外国語を音訳することである。それに対して、中国語では漢字を用いて外国語を翻字する。しかし、漢字は表意文字であるため同音異義語が存在し、音は同じでも漢字によって意味や印象が異なる。

例えば、「コカコーラ（Coca-Cola）」は中国語で「可口可楽」と表記する。この漢字列には「美味しい楽しい」という良い印象がある。しかし、「Coca-Cola」の発音に近い漢字列には「口卡口拉」もある。この漢字列には「吐き出す」という悪い印象が

あり、飲料の名称としては不適切である。そこで、外国語を中国語に翻字する場合は、音訳だけでは不十分であり、外国語の表記に使われる漢字が持つ意味や印象も考慮する必要がある。

翻字の自動化に関する既存の研究は、「狭義の翻字」と「逆翻字」に分けられる。「狭義の翻字」は音訳によって新しい言葉を生成する処理である。「逆翻字」は既存の外来語に対して、元の外国語を特定する処理である。

狭義の翻字に関して、中国語を対象とした既存の手法[3,7,8,9]は音訳をモデル化しており、印象は考慮していない。

逆翻字に関する既存の手法[2,4,5,6,10]は、音訳をモデル化して一方の言語から他方の言語に変換する点では本研究に関連する。しかし、新しい言葉を生成する訳ではないため、本研究とは目的が異なる。

本研究は、音訳と印象の両方をモデル化して、中国語を生成する翻字手法を提案する。

2. システムの構成

本システムの構成を図 1 に示す。本システムでは日本語のカタカナ語を外国語として入力させて、中国語へ翻字する。ただし、原理的には、ローマ字で表記することができれば、日本語以外の言語も入力することができる。

さらに、入力したカタカナ語に関する「印象キーワード」もユーザが入力する。ただし、印象キーワードは中国語で入力する。

カタカナ語は、音訳モデルによって音節単位で漢字列に変換される。印象キーワードは、印象モデルによって漢字に変換される。音訳モデルから複数の訳語候補が生成されるため、印象モデルで生成された漢字の確率を考慮して訳語の順位を決定する。

図 1 では、日本語「ビタミン」が入力され、音訳モデルによって、「ビタミン」と発音する漢字列（「维塔命」、「维他命」、「韦他命」など）とそれらの確率が得られている。「ビタミン」は栄養素であるため、「维护（守る）」、「他人（他人）」、「生存（生きる）」が印象キーワードとして入力されている。印象モデルに基づいて、印象キーワードに関係する漢字とそれらの確率が得られる。最後に音訳モデルと印象モデルで個別に得られた確率を統合し、最大の確率を持つ漢字列を訳語として出力する。図 1 の例では「维他命」が訳語となる。

3. 翻字ための確率モデル

日本語のカタカナ語から中国語に翻字するために、日中対訳辞書[1]に定義された発音を利用する。しかし、1 つの発音に対応する訳語は複数存在する。そこで、式(1)に示す確率を用いて訳語曖昧性を解消する。

$$
\begin{aligned}
P(K \mid R,W) &= \frac{P(R,W \mid K)\cdot P(K)}{P(R,W)} \\
&= \frac{P(R \mid K)\cdot P(W \mid K)\cdot P(K)}{P(R,W)} \\
&\propto P(R \mid K)\cdot P(W \mid K)\cdot P(K)
\end{aligned}
\tag{1}
$$

$P(K \mid R,W)$は、日本語に対するローマ字表記 R と印象キーワード W が与えられた条件のもとで中国語の漢字列 K が生成される条件付き確率である。この確率が最大の漢字列 K を訳語とする。R と W はそれぞれ独立と仮定する。$P(R,W)$ は K に依存しない定数であるため、複数の訳語候補の比較には影響しない。そこで、$P(R,W)$ は無視する。$P(R \mid K)$、$P(W \mid K)$、$P(K)$ をそれぞれ「音訳モデル」、「印象モデル」、「言語モデル」と呼ぶ。言語モデル $P(K)$ は、漢字列 K の出現確率を与える。しかし、本手法では新しい語を作るため、全ての K に対して等しい確率 $P(K)$ を与える。そこで、本手法では音訳モデルと印象モデルが重要である。

3.1 音訳モデル

音訳モデルは式(2)を用い、中国語の漢字列 K が与えられた条件のもとで日本語のローマ字表記 R が生成される条件付き確率を求める。ローマ字から漢字への変換において、ピンイン Y を中間言語として中継する。

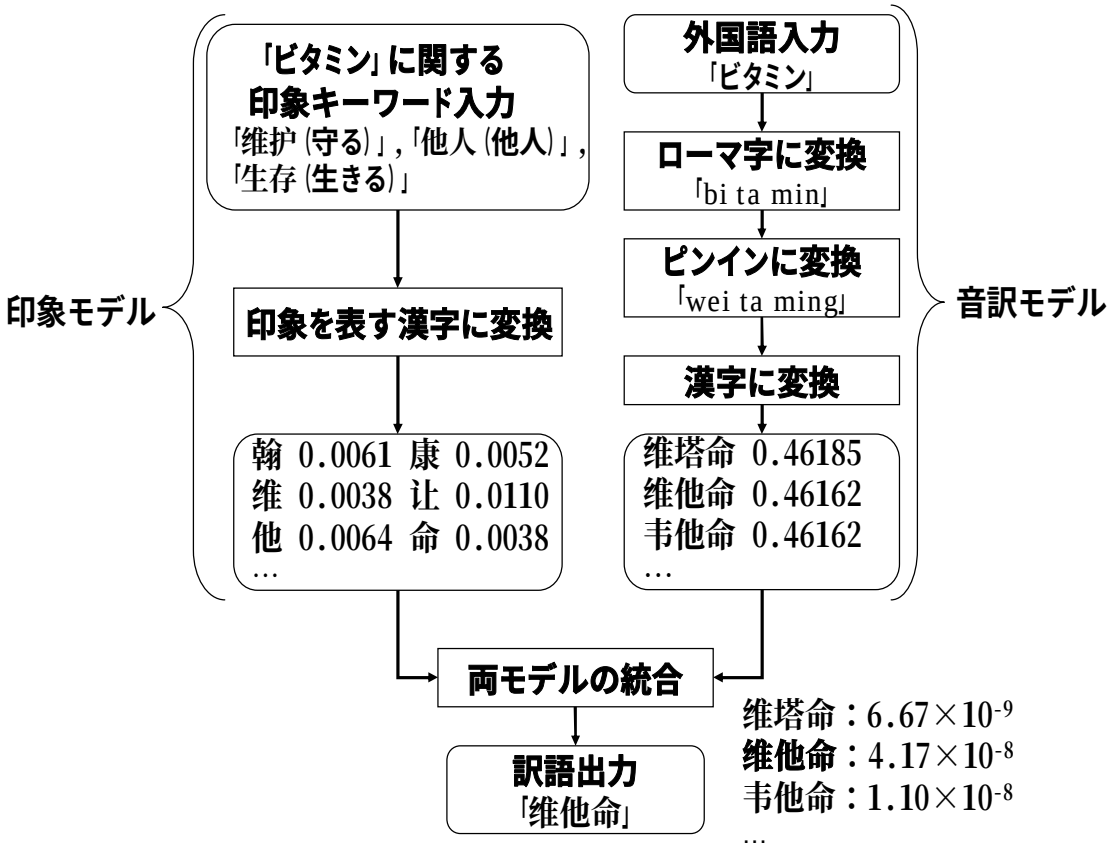

図１：言葉の印象に基づく自動翻字システムの概要

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

式(2)では、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部分列 r_i , y_i , k_i の単位で確率を計算する。

図1の例では、漢字列「维他命」が与えられた条件のもとでローマ字列「bitamin」が生成される確率を求める。具体的には、ピンイン「wei ta ming」を中継して、次のように確率を計算する。

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)

式(2)における $P(r_i \mid y_i)$ は、式(3)を用いて計算する。

$$P(r_i \mid y_i) = \frac{F(r_i, y_i)}{\sum_{r} F(r, y_i)} \qquad (3)$$

r_i と y_i はそれぞれローマ字ピンインで表記された音節である。$F(r_i, y_i)$ は、r_i と y_i がローマ字とピンインの対訳において対応する頻度である。

$F(r_i, y_i)$ を計算するために、日中対訳辞書[1]中の中国語（ピンイン付き）と対応するカタカナ語 1140 対を参考にして、ローマ字とピンインの音節を人手で対応付けた。

しかし、日本語では１つの発音が中国語では複数の発音に対応する。例えば、ローマ字の「bi」に対応するピンインは「wei」「bi」「pi」などがあるため全て考慮する。

また、日本語と中国語では音節の単位や区

切り方が異なる。そこで、可能な対応が複数存在する場合は全てを考慮する。例えば、「ビタミン」をローマ字に変換すると、「bi -ta -mi -n」と「bi -ta -min」という 2 通りのパターンが考えられるため、両方とも考慮する。

式(2)における $P(y_i \mid k_i)$ は、式(4)を用いて計算する。

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

y_i はピンインで表記された音節であり、k_i は漢字 1 文字である。$F(y_i, k_i)$ は y_i と k_i が対応する頻度である。$F(r_i, y_i)$ の計算と同じように、日中対訳辞書[1]を利用して、中国語の漢字とピンインを人手で対応付けた。

3.2 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 $P(W \mid K)$ である。

単語と漢字の対応確率を求めるために、中国語の漢字字典[11]を利用した。この漢字字典から、外来語の表記によく使われる見出し漢字（親字）599 件を用いた。見出し漢字の意味記述を形態素解析して単語を抽出し、見出し漢字と単語の共出現頻度を求めた。中国語の形態素解析には SuperMorpho（オムロンソフト社）を用いた。式(5)を用いて $P(W \mid K)$ を計算する。

$$P(W \mid K) = \prod_i P(w_i \mid k_i) \qquad (5)$$

w_i は W を構成する単語 1 つであり、k_i は K を構成する漢字 1 文字である。$P(w_i \mid k_i)$ は式(6)を用いて計算する。

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

k_i は字典の見出し漢字である。w_i は k_i の意味記述に現れる単語である。$F(k_i \mid w_i)$ は w_i が k_i の意味記述に使用された頻度である。

3.3 音訳モデルと印象モデルの統合

本手法の最終目的として、音訳モデルと印象モデルの結果を統合する。

図 2 は図 1 を簡略化したものであり、左側は印象モデルから得られた漢字とそれらの確率である。右側は音訳モデルから得られた「ビタミン」の発音に近い漢字列とそれらの確率である。そして、音訳モデルで得られた漢字と印象モデルで得られた漢字を照合する。音訳モデルの漢字と印象モデルの漢字が一致した場合、式(1)に示したように、2 つの確率を掛け合わせる。

ただし、一致しない漢字には確率が計算できないため、特定の定数を与える。すなわち、一致しない漢字のために漢字列全体の確率がゼロにならないよう平滑化を行う。この定数は経験的に 0.001 とした。最後に、一番高い確率を持つ漢字列を訳語として出力する。

4. 評価実験

本手法の有効性を評価するため、日本語のカタカナ語に対して、音訳モデルだけで翻字した結果と、音訳モデルと印象モデルを統合した本手法の結果を比較した。評価用の日本語として、日中対訳辞書[1]に登録されたカタカナ語 1140 語のうち、対訳の中国語と発音が似ている 210 語を選び、評価に使用した。210 語の内訳は、商品名（63 件）、企業名（52 件）、地名（42 件）、一般名詞（32 件）、人名（21 件）であった。

日本語が分かる中国人に判定を依頼し、評価対象のカタカナ語に関する印象キーワードを入力してもらった。なお、判定者には評価対象のカタカナ語について理解できるような説明を与えた。

次に、そのカタカナ語について音訳モデルと本手法で個別に翻字を行い、訳語（中国語の漢字列）の順位付きリストを生成した。

対訳辞書に定義された訳語以外にも適切な訳語が存在する可能性がある。そこで、判定者には適切と判断した訳語を網羅的に特定してもらった。しかし、2 つのリストを個別に判定してもらうと、判定するリストの順番によって判定結果が変わる可能性がある。そこで、2 つのリストから上位 100 語ずつを抽出して併合し、重複する語を除いて文字コード

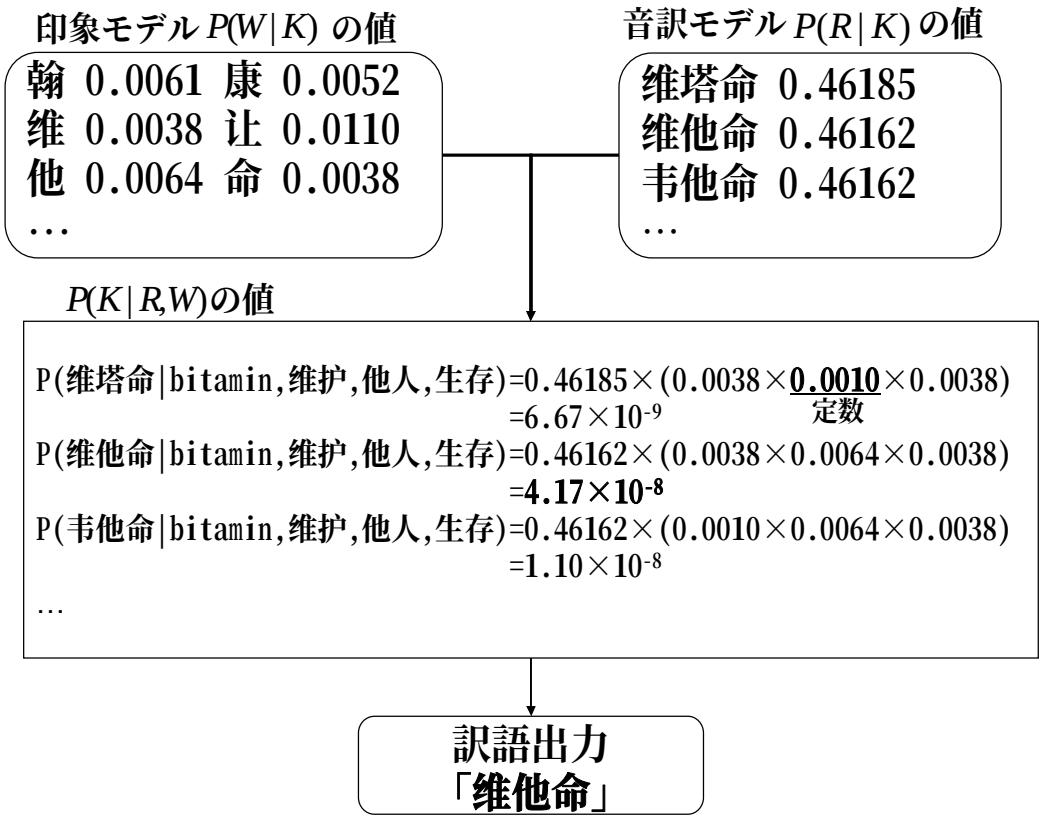

図 2：音訳モデルと印象モデルの統合

でソートした。こうすることで、判定者には、ないようにした。判定者には、併合したリスト中の語に対して、自分が入力した印象を考慮しながら、訳語としての適否を判定してもらった。

しかし、以上の判定は主観的であるため、3 人の判定者に同じ 210 語に対する評価を個別に依頼した。評価に使用したカタカナ語の例を表 1 に示す。表 1 では、カタカナ語の対訳である中国語と各判定者が与えた印象キーワードも示している。

評価尺度として、各手法が生成したリストにおける「正解訳語の平均順位」を用いた。1 つのカタカナ語に対して複数の正解訳語がある場合は、まずそれらの順位を平均し、さらに全カタカナ語を横断して順位を平均した。

「正解」として、以下に示す 3 通りの解釈が存在する。

① 判定者が個別に正解と判定した訳語
② 判定者全員が正解と判定した訳語
③ 日中対訳辞書に登録されている訳語

上記①を正解とした場合の評価結果を表 2 に示す。どの判定者の結果でも、本手法は音訳モデルだけの手法に比べて正解の平均順位を向上させた。また、判定者 3 名の結果を平均すると、音訳モデルでは正解の平均順位が 731 位だった。それに対して、本手法では正解の平均順位が 51 位だった。

上記②の正解は、判定者全員の意見が一致した場合であり、①に比べると客観性が高い。1 つのカタカナ語に複数の正解が判定の正解が判定された場合は、それら全てが一致したカタカナ語だけを評価の対象とした。

どの訳語がどの手法で生成されたのか分から

表 1：評価実験に使用したカタカナ語、辞書に定義された中国語の対訳、判定者が与えた印象キーワードの例

カタカナ語（種別）	中国語	印象キーワード（括弧内は日本語訳）		
		判定者 A	判定者 B	判定者 C
アウディ（商品名）	奥迪	轿车(乘用車),富贵(富貴),品质(品質),速度(速度)	车名(車名),德国(ドイツ),豪华(豪華),气派(気概),价格(価格)	高贵(高い),速度(速度),德国(ドイツ)
エプソン（企業名）	爱普生	印刷机(印刷機),知名(知名),品质(品質),优异(優秀)	电脑(計算機),打印机(印刷機),公司(会社),产品(製品),日本(日本)	喜爱(好み),普及(普及),生动(生動)
エンジェル(一般名詞)	安琪儿	天使(天使),可爱(可愛い),幸福(幸福),爱心(愛)	神话(神話),浪漫(ロマンチック),天空(空),白色(白い),美丽(綺麗)	天使(天使),平安(平安),可爱(可愛い),儿童(子供)
カネボウ（企業名）	嘉娜宝	美丽(美しい),化妆品(化粧品),肌肤(皮膚),女人(女)	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(よし),宝贝(宝物)
シャープ（企業名）	夏普	电器(電気),普及(普及),省电(省エネルギー)	电子产品(電子製品),日本(日本),种类(種類)	普及(普及),电器(電気),夏天(夏)
ショパン（人名）	肖邦	月光(月光),钢琴(ピアノ),古典(古典),音乐(音楽),欣赏(鑑賞)	人名(人名),音乐(音楽),作曲(作曲),演奏(演奏)	美好(良い),邦国(外国),国家(国)

②の場合は、判定者によって正解は変わらない。しかし、判定者によって与えられた印象キーワードは異なるため、本手法で生成されたリストにおける訳語の順位は変わる。②を正解とした場合の実験結果を表 3 に示す。判定者 3 人の判定が一致したカタカナ語は 108 語あり、正解訳語は 120 語あった。判定者によらず、本手法は音訳モデルだけの手法に比べて正解の平均順位を向上させた。

上記③に対する結果を表 4 に示す。表 2 や表 3 と同じように、本手法は、判定者によらず、音訳モデルだけの手法に比べて正解の平均順位を向上させた。

表 2〜4 の結果より、本手法で提案した印象モデルの有効性が確認された。

さらに、判定者が与えた印象キーワードの数によって正解の平均順位がどのように変化するかを調べた。

表 2:正解①に対する実験結果

判定者	カタカナ語数	正解訳語数	正解の平均順位	
			音訳	音訳＋印象
A	192	758	511	84
B		716	652	43
C		492	1021	28
平均	——	655	731	51

表 3：正解②に対する実験結果

判定者	カタカナ語数	正解訳語数	正解の平均順位	
			音訳	音訳＋印象
A	108	120	283	22
B				23
C				18
平均	——	——	283	21

表 4：正解③に対する実験結果

判定者	カタカナ語数	正解訳語数	正解の平均順位	
			音訳	音訳＋印象
A	210	210	1738	260
B				249
C				103
平均	——	——	1738	204

判定者には、適切さに基づいて印象キーワードに順位を付けてもらった。そこで、順位が高い方から印象キーワードの数を徐々に増やした。正解訳語として①の解釈を使用した。実験結果を表 5 に示す。入力された印象キーワード数が多いほど、正解訳語の平均順位が高くなった。

表 5：印象キーワード数と正解の平均順位

判定者	印象キーワード数		
	1	2	3
A	101	94	93
B	62	58	52
C	101	70	33
平均	88	74	59

表 2 において、印象モデルのために正解訳語の順位が下がった原因について分析した。判定者全員の正解訳語総数は 1966 語あり、そのうち 639 語は印象モデルの統合によって正解の順位が下がった。その主な原因は、印象モデルによって生成された漢字が適切でなかったことであった。そこで、今後は印象モデルに含まれる漢字や単語を洗練する必要がある。

5. まとめ

本研究は発音と印象の両方を考慮して外国語を中国語に翻字する手法を提案した。また、評価実験によって本手法の有効性を示した。今後の研究課題は、評価実験の誤り分析を通して手法を洗練することである。

参考文献：

[1] 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 年

[2] Chen, H. H., S.J.Huang, Y.W.Ding, and S.C.Tsai. “Proper Name Translation in Cross-Language Information Retrieval”. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics, pp.232-236, 1998.

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

[4] Knight K. and J. Graehl. “Machine Transliteration”. Computational Linguistics, Vol.24, No.4, 599-612, 1998.

[5] 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.

[6] Kil Soon Jeong, Sung Hyon Myaeng, Jae Sung Lee, and Key-Sun Choi. “Automatic identification and back-transliteration of foreign words for information retrieval”. Information Processing and Management, pp.523-540, 1999.

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

[9] Chun-Jen 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.

[10] Yan Qu and Gregory Grefenstette. “Finding Ideographic Representations of Japanese Names Written in Latin Script via Language Identification and Corpus Validation”. Proceedings of the 42nd Meeting of the Association for Computational Linguistics, pp. 183-190, 2004.

[11] 新華字典電子版 1.0