Modeling Impression in Probabilistic Transliteration into Chinese

LiLi Xu* **Atsushi Fujii**
Graduate School of Library,
Information and Media Studies
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
fujii@slis.tsukuba.ac.jp

Tetsuya Ishikawa
The Historiographical Institute
The University of Tokyo
3-1 Hongo 7-chome, Bunkyo-ku
Tokyo, 133-0033, Japan
ishikawa@hi.u-tokyo.ac.jp

Abstract

For transliterating foreign words into Chinese, the pronunciation of a source word is spelled out with Kanji characters. Because Kanji comprises ideograms, an individual pronunciation may be represented by more than one character. However, because different Kanji characters convey different meanings and impressions, characters must be selected carefully. In this paper, we propose a transliteration method that models both pronunciation and impression, whereas existing methods do not model impression. Given a source word and impression keywords related to the source word, our method derives possible transliteration candidates and sorts them according to their probability. We evaluate our method experimentally.

1 Introduction

Reflecting the rapid growth of science, technology, and economies, new technical terms and product names have progressively been created. These new words have also been imported into different languages. There are three fundamental methods for importing foreign words into a language.

In the first method—*translation*—the meaning of the source word in question is represented by an existing or new word in the target language.

In the second method—*transliteration*—the pronunciation of the source word is represented by using the phonetic alphabet of the target language, such as Katakana in Japanese and Hangul in Korean.

In the third method, the source word is spelled out as it is. However, the misuse of this method decreases the understandability and readability of the target language.

While translation is time-consuming, requiring selection of an existing word or generation of a new word that correctly represents the meaning of the source word, transliteration can be performed rapidly. However, the situation is complicated for Chinese, where a phonetic alphabet is not used and Kanji is used to spell out both conventional Chinese words and foreign words.

Because Kanji comprises ideograms, an individual pronunciation can potentially be represented by more than one character. However, if several Kanji strings are related to the same pronunciation of the source word, their meanings will be different and will therefore convey different impressions.

For example, “Coca-Cola” can be represented by different Kanji strings in Chinese. The official transliteration is “可口可乐”, which comprises “可口 (tasty)” and “可乐 (pleasant)”, and is therefore associated with a positive connotation.

However, there are a number of Kanji strings that represent similar pronunciations to that of “Coca-Cola”, but which are associated with inappropriate impressions for a beverage, such as “口卡口拉”. This word includes “口卡”, which is associated with choking.

Therefore, Kanji characters must be selected carefully during transliteration into Chinese. This is especially important when foreign companies intend to introduce their names and products into China.

In this paper, we propose a method that models both impression and pronunciation for transliteration into Chinese.

* This work was done when the first author was a graduate student at University of Tsukuba, who currently works for Hitachi Construction Machinery Co., Ltd.

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 242–249,
Sydney, July 2006.

Section 2 surveys previous research into automatic transliteration, in order to clarify the meaning and contribution of our research. Section 3 elaborates on our transliteration method. Section 4 evaluates the effectiveness of our method.

2 Related Work

In a broad sense, the term "transliteration" has been used to refer to two tasks.

The first task is transliteration in the strict sense, which creates new words in a target language (Haizhou et al., 2004; Wan and Verspoor, 1998).

The second task is back-transliteration (Fujii and Ishikawa, 2001; Jeong et al., 1999; Knight and Graehl, 1998; Qu et al., 2003), which identifies the source word corresponding to an existing transliterated word. Back-transliteration is intended mainly for cross-lingual information retrieval and machine translation.

Both transliteration tasks require methods that model pronunciation in the source and target languages.

However, by definition, in back-transliteration, the word in question has already been transliterated and the meaning or impression of the source word does not have to be considered. Thus, back-transliteration is outside the scope of this paper.

In the following, we use the term "transliteration" to refer to transliteration in the strict sense.

Existing transliteration methods for Chinese (Haizhou et al., 2004; Wan and Verspoor, 1998) aim to spell out foreign names of people and places, and do not model impression.

However, as exemplified by "Coca-Cola" in Section 1, the impression of words needs to be modeled in the transliteration of proper names, such as companies and products. The contribution of our research is to incorporate a model of impression into automatic transliteration.

3 Methodology

3.1 Overview

Figure 1 shows our transliteration method, which models both pronunciation and impression when transliterating foreign words into Chinese. We will explain the entire process of our transliteration method in terms of Figure 1.

The input for our method is twofold. First, a source word to be transliterated into Chinese is requested. Second, one or more words that describe

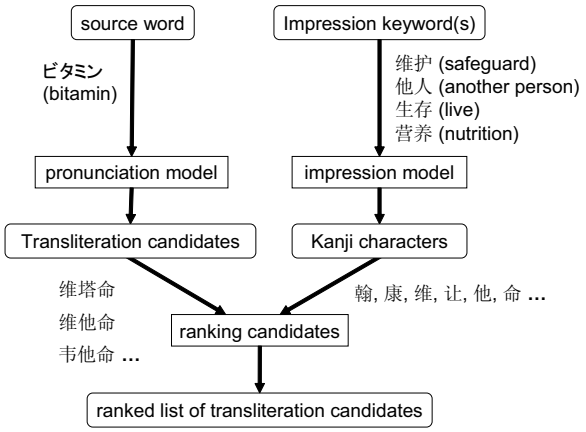

Figure 1: Overview of our transliteration method for Chinese.

the impression of the source word, which we call "impression keywords", are requested. Currently, impression keywords must be provided manually in Chinese. The output of our method is one or more Kanji strings.

In an example scenario using our method, a user has a good command of Chinese and intends to introduce something (e.g., a company or product) into China. It is reasonable to assume that this user can provide one or more Chinese impression keywords to associate with the target object.

Using the pronunciation model, the source word is converted into a set of Kanji strings whose pronunciation is similar to that of the source word. Each of these Kanji strings is a transliteration candidate.

Currently, we use Japanese Katakana words as source words, because Katakana words can easily be converted into pronunciations using the Latin alphabet. However, in principle, any language that uses phonetic script can be a source language for our method. In Figure 1, the Katakana word "*bitamin* (vitamin)" is used as an example source word.

Using the impression model, impression keywords are converted into a set of Kanji characters. A simple implementation is to segment each impression keyword into characters.

However, because it is difficult for a user to provide an exhaustive list of appropriate keywords and characters, our impression model derives characters that are not included in the impression keywords.

Because of the potentially large number of selected candidates, we need to rank the candidates. We model both pronunciation and impression in

a probabilistic framework, so that transliteration candidates are sorted according to their probability score.

Transliteration candidates that include many characters derived from the impression model are preferred. In other words, the Kanji characters derived via the impression model are used to rerank the candidates derived via the pronunciation model.

We elaborate on our probabilistic transliteration model in Section 3.2. We then discuss the pronunciation and impression models in Sections 3.3 and 3.4, respectively.

3.2 Probabilistic Transliteration Model

Given a romanized Japanese Katakana word R and a set of impression keywords W, our purpose is to select the Kanji string K that maximizes $P(K|R,W)$, which is evaluated as shown in Equation (1), using Bayes' theorem.

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

In the second line of Equation (1), we assume the conditional independence of R and W given K. In the third line, we omit $P(R,W)$, which is independent of K. This does not affect the relative rank of Kanji strings, when ranked in terms of $P(K|R,W)$.

In Figure 1, R and W are "bitamin" and "维护 他人 生存 营养", respectively, and a K candidate is "维塔命".

If a user intends to select more than one Kanji string, those Ks associated with higher probabilities should be selected.

As shown in Equation (1), $P(K|R,W)$ can be approximated by the product of $P(R|K)$, $P(W|K)$, and $P(K)$. We call these three factors the pronunciation, impression, and language models, respectively.

The language model, $P(K)$, models the probability of K irrespective of R and W. In probabilistic natural language processing, $P(K)$ is usually realized by a word or character N-gram model, and therefore a K that appears frequently in a corpus is assigned a high probability.

However, because our purpose is to generate new words, the use of statistics obtained from existing corpora is not effective. Therefore, we consider $P(K)$ to be constant for every K.

In summary, $P(K|R,W)$ is approximated by a product of $P(R|K)$ and $P(W|K)$. The quality of our transliteration method will depend on the implementation of the pronunciation and impression models.

3.3 Pronunciation Model

The pronunciation model, $P(R|K)$, models the probability that a roman representation R is selected, given a Kanji string K.

In Japanese, the Hepburn and *Kunrei* systems are commonly used for romanization purposes. We use the Hepburn system. We use Pinyin as a representation for Kanji characters. We decompose K into Kanji characters and associate K with R on a character-by-character basis. We calculate $P(R|K)$ as shown in Equation (2).

$$\begin{aligned} P(R|K) &\approx P(R|Y)\cdot P(Y|K) \\ &\approx \prod_{i=1}^{N} P(r_i|y_i)\cdot \prod_{j=1}^{N} P(y_j|k_j) \end{aligned} \tag{2}$$

Y denotes the Pinyin strings representing the pronunciation of K. k_i denotes a single Kanji character. r_i and y_i denote substrings of R and Y, respectively. R, Y, and K are decomposed into the same number of elements, namely N. We calculate $P(r_i|y_i)$ and $P(y_i|k_i)$ as shown in Equation (3).

$$\begin{aligned} P(r_i|y_i) &= \frac{F(r_i,y_i)}{\sum_r F(r,y_i)} \\ P(y_i|k_i) &= \frac{F(y_i,k_i)}{\sum_y F(y,k_i)} \end{aligned} \tag{3}$$

$F(x,y)$ denotes the co-occurrence frequency of x and y. We need the co-occurrence frequencies of r_i and y_i and the co-occurrence frequencies of y_i and k_i in order to calculate $P(R|K)$.

We used a bilingual dictionary comprising 1 140 Katakana words, most of which are technical terms and proper nouns, and their transliterations into Chinese, which are annotated with Pinyin. We manually corresponded 151 pairs of Katakana and roman characters on a mora-by-mora basis, and romanized Katakana characters in the dictionary automatically.

We obtained 1 140 tuples, of the form $< R,Y,K >$. Because the number of tuples was

manageable, we obtained the element-by-element R, Y, and K correspondences manually. Finally, we calculated $F(r_i, y_i)$ and $F(y_i, k_i)$.

If there are many tuples, and the process of manual correspondence is expensive, we can automate the process as performed in existing transliteration methods, such as the EM algorithm (Knight and Graehl, 1998) or DP matching (Fujii and Ishikawa, 2001).

The above calculations are performed off-line. In the online process, we consider all possible segmentations of a single Katakana word. For example, the romanized Katakana word "*bitamin* (vitamin)" corresponds to two Pinyin strings and is segmented differently, as follows:

- bi-ta-min: wei-ta-ming,
- bi-ta-mi-n: wei-ta-mi-an.

3.4 Impression Model

The impression model, $P(W|K)$, models the probability that W is selected as a set of impression keywords, given Kanji string K. As in the calculation of $P(R|K)$ in Equation (2), we decompose W and K into elements, in calculating $P(W|K)$.

W is decomposed into a set of words, w_i, and K is decomposed into a set of Kanji characters, k_j. We calculate $P(W|K)$ as a product of $P(w_i|k_j)$, which is the probability that w_i is selected as an impression keyword given k_j.

However, unlike Equation (2), the numbers of w_i and k_j derived from W and K are not always the same, because users are allowed to provide an arbitrary number of impression keywords. Therefore, for each k_j we select the w_i that maximizes $P(w_i|k_j)$ and approximate $P(W|K)$ as shown in Equation (4).

$$P(W|K) \approx \prod_j \max_{w_i} P(w_i|k_j) \qquad (4)$$

Figure 2 shows an example in which the four Chinese words in the "w_i" column are also used in Figure 1.

We calculate $P(w_i|k_j)$ by Equation (5).

$$P(w_i|k_j) = \frac{F(w_i, k_j)}{\sum_w F(w, k_j)} \qquad (5)$$

As in Equation (3), $F(x, y)$ denotes the co-occurrence frequency of x and y.

w_i \ k_j	维	他	命
维护	0.5	——	——
他人	0.3	0.4	——
生存	——	——	0.6
营养	0.1	——	——

P(维护 他人 生存 营养|维他命)
= P(维护|维)×P(他人|他)×P(生存|命)
= 0.5×0.4×0.6

Figure 2: Example calculation of $P(W|K)$.

In summary, we need co-occurrences of each word and character in Chinese.

These co-occurrences can potentially be collected from existing language resources, such as corpora in Chinese.

However, it is desirable to collect an *association* between a word and a character, not simply their co-occurrence in corpora. Therefore, we used a dictionary of Kanji in Chinese, in which each Kanji character entry is explained via sentences, and often exemplified by one or more words that include that character.

We selected 599 entry characters that are often used to spell out foreign words. Then we collected the frequencies with which each word is used to explain each entry character.

Because Chinese sentences lack lexical segmentation, we used SuperMorpho[1] to perform a morphological analysis of explanation sentences and example words. As a result, 16 943 word types were extracted. We used all of these words to calculate the co-occurrence frequencies, irrespective of the parts of speech.

Table 1 shows examples of Kanji characters, Chinese words, and their co-occurrence frequencies in the dictionary.

However, $P(w_i|k_j)$ cannot be calculated for the Kanji characters not modeled in our method (i.e., the Kanji characters not included in the 599 entry characters). Thus, for smoothing purposes, we experimentally set $P(w_i|k_j)$ at 0.001 for those k_j not modeled.

4 Experiments

4.1 Method

We evaluated our transliteration method experimentally. Because the contribution of our research is the incorporation of the impression model in a transliteration method, we used a method that uses only the pronunciation model as a control.

[1] http://www.omronsoft.com/

Table 1: Example of characters, words, and their co-occurrence frequencies.

k_j	w_i	$F(w_i,k_j)$	k_j	w_i	$F(w_i,k_j)$	k_j	w_i	$F(w_i,k_j)$
高	高	39	好	美	3	乐	喜悦	2
高	高大	8	好	貌美	2	乐	愉快	1
高	远	4	好	好	43	乐	快乐	5
高	下	4	好	好看	2	乐	幸福	2
高	距离	2	好	美丽	2	乐	乐	51
高	大	1	好	好不	2	乐	笑	5
高	俗	2	好	好吃	2	乐	喜	3
高	崇高	2	好	表示	4	乐	音乐	11
高	高尚	2	好	同意	2	乐	乐意	2
高	加高	3	好	喜好	1	乐	安乐	7
高	增高	1	好	喜爱	2	乐	乐于	5

From a Japanese–Chinese dictionary, we selected 210 Katakana words that had been transliterated into Chinese, and used these Katakana words as test words. Each test word can be classified into one of the following five categories: products, companies, places, persons, or general words. Details of the categories of test inputs are shown in Table 2.

Three Chinese graduate students who had a good command of Japanese served as assessors and produced reference data. None of the assessors was an author of this paper. The assessors performed the same task for the same test words independently, in order to enhance the objectivity of the results.

We produced the reference data via the following procedure.

First, for each test word, each assessor provided one or more impression keywords in Chinese. We did not restrict the number of impression keywords per test word, which was determined by each assessor.

If an assessor provided more than one impression keyword for a single test word, he/she was requested to sort them in order of preference, so that we could investigate the effect of the number of impression keywords on the evaluation results, by changing the number of top keywords used for transliteration purposes.

We provided the assessors with the descriptions for the test words from the source dictionary, so that the assessors could understand the meaning of each test word.

Second, for each test word, we applied the control method and our method independently, which produced two lists of ranked transliteration candidates. Because the impression keywords provided by the assessors were used only in our method, the ranked list produced by the control was the same for all assessors.

Table 2: Categories of test words.

Category	# Words	Example word		
		Japanese	Chinese	English
Product	63	アウディ	奥迪	Audi
Company	49	エプソン	爱普生	Epson
Place	36	オハイオ	俄亥俄	Ohio
Person	21	ショパン	肖邦	Chopin
General	41	エンジェル	安琪儿	angel

Third, for each test word, each assessor identified one or more correct transliterations, according to their impression of the test word. It was important not to reveal to the assessors which method produced which candidates.

By these means, we selected the top 100 transliteration candidates from the two ranked lists for the control and our method. We merged these candidates, removed duplications, and sorted the remaining candidates by the character code.

As a result, the assessors judged the correctness of up to 200 candidates for each test word. However, for some test words, assessors were not able to find correct transliterations in the candidate list.

The resultant reference data was used to evaluate the accuracy of a test method in ranking transliteration candidates. We used the average rank of correct answers in the list as the evaluation measure. If more than one correct answer was found for a single test word, we first averaged the ranks of these answers and then averaged the ranks over the test words.

Although we used the top 100 candidates for judgment purposes, the entire ranked list was used to evaluate each method. Therefore, the average rank of correct answers can potentially be over 100. The average number of candidates per test word was 31 779.

Because our method uses the impression model to re-rank the candidates produced by the pronunciation model, the lists for the control and our method comprise the same candidates. Therefore, it is fair to compare these two methods by the average rank of the correct answers.

For each test word, there is more than one type of “correct answer”, as follows:

(a) transliteration candidates judged as correct by the assessors independently (translitera-

tion candidates judged as correct by at least one assessor);

(b) transliteration candidates judged as correct by all assessors;

(c) transliterations defined in the source dictionary.

In (a), the coverage of correct answers is the largest, whereas the objectivity of the judgment is the lowest.

In (c), the objectivity of the judgment is the largest, whereas the coverage of correct answers is the lowest. Although for each Katakana word the source dictionary gives only one transliteration that is commonly used, there are a number of appropriate out-of-dictionary transliterations.

In (b), where the assessors did not disagree about the correctness, the coverage of correctness and the objectivity are both middle ranked.

Because none of the above answer types is perfect, we used all three types independently.

4.2 Results and Analyses

Tables 3–5 show the results of comparative experiments using the answer types (a)–(c) above, respectively.

In Tables 3–5, the column "# of test words" denotes the number of test words for which at least one correct answer exists. While the values in the second column of Table 3 are different depending on the assessor, in Tables 4 and 5 the values of the second column are the same for all assessors.

The columns "Avg. # of KW" and "Avg. # of answers" denote the number of impression keywords and the number of correct answers per test word, respectively. While the values in the fourth column of Table 3 are different depending on the assessor, in Tables 4 and 5 the values of the fourth column are the same for all assessors.

In Tables 4 and 5, the average rank of correct answers for the control is the same for all assessors. However, the average rank of correct answers for our method is different depending on the assessor, because the impression keywords used depended on the assessor.

The two columns in "Avg. rank" denote the average ranks of correct answers for the control and for our method, respectively. Looking at Tables 3–5, it can be seen that our method outperformed the control in ranking transliteration candidates, irrespective of the assessor and the answer type.

The average rank of correct answers for our method in Table 5 was lower than those in Tables 3 and 4. One reason is that the correct answers in the source dictionary are not always related to the impression keywords provided by the assessors.

Table 6 presents the results in Table 3 on a category-by-category basis. Because the results were similar for answer types (b) and (c), we show only the answer type (a) results, for the sake of conciseness. Looking at Table 6, it can be seen that our method outperformed the control in ranking transliteration candidates, irrespective of the category of test words.

Our method was effective for transliterating names of places and people, although these types of words are usually transliterated independently of their impressions, compared with the names of products and companies.

One reason is that, in the dictionary of Kanji used to produce the impression model, the explanation of an entry sometimes includes a phrase, such as "this character is often used for a person's name". Assessors provided the word "person" in Chinese as an impression keyword for a number of person names. As a result, transliteration candidates that included characters typically used for a person's name were highly ranked.

It may be argued that, because the impression model was produced using Kanji characters that are often used for transliteration purposes, the impression model could possibly rank correct answers better than the pronunciation model. However, the pronunciation model was also produced from Kanji characters used for transliteration purposes.

Figure 3 shows the distribution of correct answers for different ranges of ranks, using answer type (a). The number of correct answers in the top 10 for our method is approximately twice that of the control. In addition, by our method, most of the correct answers can be found in the top 100 candidates. Because the results were similar for answer types (b) and (c), we show only the answer type (a) results, for the sake of conciseness.

As explained in Section 4.1, for each test word, the assessors were requested to sort the impression keywords in order of preference. We analyzed the relation between the number of impression keywords used for the transliteration and the average rank of correct answers, by varying the threshold for the number of top impression keywords used.

Table 3: Results obtained with answer type (a).

Assessor	# of test words	Avg. # of KW	Avg. # of answers	Avg. rank	
				Control	Our method
A	205	5.1	3.8	706	82
B	204	5.8	3.8	728	44
C	199	3.5	2.6	1 130	28
Avg.	203	4.8	3.4	855	51

Table 4: Results obtained with answer type (b).

Assessor	# of test words	Avg. # of KW	Avg. # of answers	Avg. rank	
				Control	Our method
A	108	5.1	1.1	297	22
B	108	5.8	1.1	297	23
C	108	3.5	1.1	297	18
Avg.	108	4.8	1.1	297	21

Table 5: Results obtained with answer type (c).

Assessor	# of test words	Avg. # of KW	Avg. # of answers	Avg. rank	
				Control	Our method
A	210	5.1	1	1 738	260
B	210	5.8	1	1 738	249
C	210	3.5	1	1 738	103
Avg.	210	4.8	1	1 738	204

Table 6: Results obtained with answer type (a) on a category-by-category basis.

Category	# of test words	Avg. # of KW	Avg. # of answers	Avg. rank	
				Control	Our method
Product	144	4.8	3.5	1 527	64
Company	186	4.7	3.6	742	54
Place	102	4.8	3.7	777	46
Person	61	5.0	3.4	766	51
General	115	4.7	2.6	280	38
Avg.	122	4.8	3.4	818	51

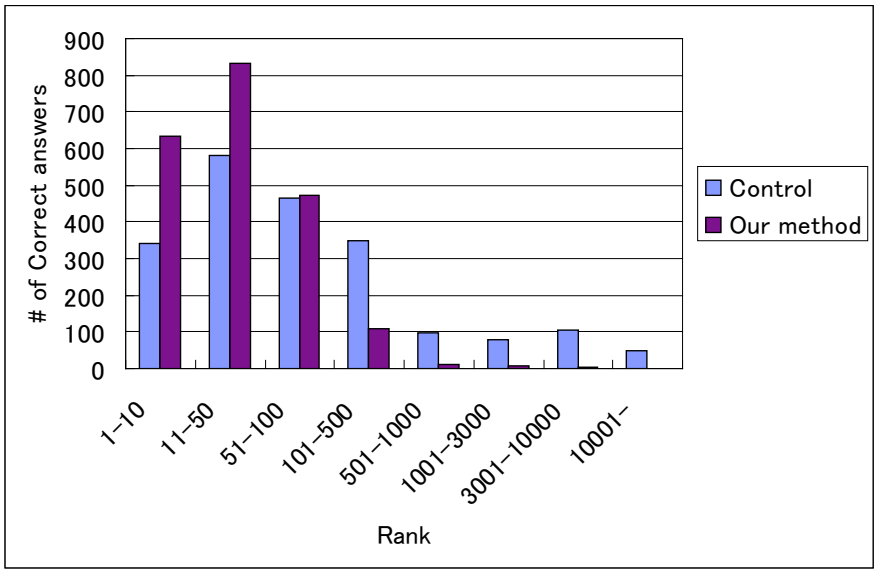

Figure 3: Distribution of average rank for correct answers.

Table 7 shows the average rank of correct answers for different numbers of impression keywords, on an assessor-by-assessor basis. By comparing Tables 3 and 7, we see that even if a single impression keyword was provided, the average rank of correct answers was higher than that for the control. In addition, the average rank of correct answers was generally improved by increasing the number of impression keywords.

Finally, we investigated changes in the rank of correct answers caused by our method. Table 8 shows the results, in which "Higher" and "Lower" denote the number of correct answers whose ranks determined by our method were higher or lower, respectively, than those determined by the control.

For approximately 30% of the correct answers, our method decreased the control's rank. Errors were mainly caused by correct answers containing Kanji characters that were not modeled in the impression model. Although we used a smoothing technique for characters not in the model, the result was not satisfactory. To resolve this problem, the number of characters in the impression model should be increased.

In summary, our method, which uses both the impression and pronunciation models, ranked correct transliterations more highly than a method that used only the pronunciation model. We conclude that the impression model is effective for transliterating foreign words into Chinese. At the same time, we concede that there is room for improvement in the impression model.

5 Conclusion

For transliterating foreign words into Chinese, the pronunciation of a source word is spelled out with Kanji characters. Because Kanji characters are ideograms, a single pronunciation can be represented by more than one character. However, because different Kanji characters convey different meanings and impressions, characters must be selected carefully.

In this paper, we proposed a transliteration method that models both pronunciation and impression, compared to existing methods that do not model impression. Given a source word and impression keywords related to the source word, our method derives possible transliteration candidates, and sorts them according to their probability. We showed the effectiveness of our method experimentally.

Table 7: Relation between the number of impression keywords and average rank of correct answers with answer type (a).

	# of KW		
Assessor	1	2	3
A	103	94	92
B	64	60	52
C	113	73	34

Table 8: Changes in ranks of correct answers caused by our method.

		Avg. rank	
Answer type	# of answers	Higher	Lower
(a)	2 070	1 431	639
(b)	360	250	110
(c)	630	422	208

Future work will include collecting impression keywords automatically, and adapting the language model to the category of source words.

References

Atsushi Fujii and Tetsuya Ishikawa. 2001. Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, 35(4):389–420.

Li Haizhou, Zhang Min, and Su Jian. 2004. A joint source-channel model for machine transliteration. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Kil Soon Jeong, Sung Hyon Myaeng, Jae Sung Lee, and Key-Sun Choi. 1999. Automatic identification and back-transliteration of foreign words for information retrieval. *Information Processing & Management*, 35:523–540.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Yan Qu, Gregory Grefenstette, and David A. Evans. 2003. Automatic transliteration for Japanese-to-English text retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 353–360.

Stephen Wan and Cornelia Maria Verspoor. 1998. 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*, pages 1352–1356.