Effects of Related Term Extraction in Transliteration into Chinese

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

Abstract

To transliterate foreign words, in Japanese and Korean, phonograms, such as Katakana and Hangul, are used. In Chinese, the pronunciation of a source word is spelled out using Kanji characters. Because Kanji is ideogrammatic representation, different Kanji characters are associated with the same pronunciation, but can potentially convey different meanings and impressions. To select appropriate Kanji characters, an existing method requests the user to provide one or more related terms for a source word, which is time-consuming and expensive. In this paper, to reduce this human effort, we use the World Wide Web to extract related terms for source words. We show the effectiveness of 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 two 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. Technical terms and proper nouns are often transliterated.

In Chinese, Kanji is used to spell out both conventional Chinese words and foreign words. Because Kanji is ideogrammatic, an individual pronunciation can be represented by more than one character. If several Kanji strings are related to the same pronunciation of the source word, their meanings will be different and convey different impressions.

For example, “Coca-Cola” can be represented by different Kanji strings in Chinese with similar pronunciations, such as “可口可乐” and “口卡口拉”. 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, “口卡口拉” is associated with a negative connotation because this word includes “口卡 ”, which is associated with “choking”.

For another example, the official transliteration of the musician Chopin’s name in Chinese is “肖邦”, where “肖” is commonly used for Chinese family names. Other Kanji characters with the same pronunciation as “肖” include “消”. However, “消”, which means “to disappear”, is not ideal for a person’s name.

Thus, 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 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; Xu et al., 2006). The second task is back-transliteration (Knight and Graehl, 1998), which identifies the source word correspond-

ing to an existing transliterated word. Both 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), which aim to spell out foreign names of people and places, do not model the impression the transliterated word might have on the reader.

Xu et al. (2006) proposed a method to model both the impression and the pronunciation for transliteration into Chinese. In this method, impression keywords that are related to the source word are used. However, a user must provide impression keywords, which is time-consuming and expensive.

In this paper, to reduce the amount of human effort, we propose a method that uses the World Wide Web to extract related terms for source words.

2 Overview

Figure 1 shows our transliteration method, which models pronunciation, impression, and target language when transliterating foreign words into Chinese. Figure 1 is an extension of the method proposed by Xu et al. (2006) and the part surrounded by a dotted line is the scheme we propose in this paper. We will explain the entire process using Figure 1.

There are two parts to the input for our method. First, a source word to be transliterated into Chinese is requested. Second, the category of the source word, such as "company" or "person", is requested. The output is one or more Kanji strings.

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 be easily converted into pronunciations using the Latin alphabet. In Figure 1, the Katakana word "epuson (EPSON)" is used as an example source word. However, in principle, any language that uses a phonetic script can be a source language for our method.

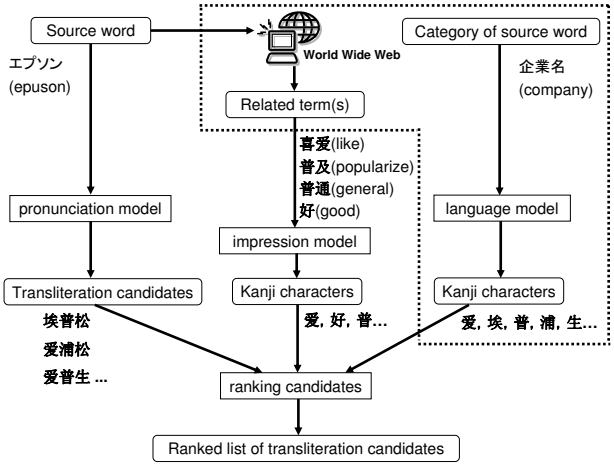

Figure 1: Overview of our transliteration method.

Using the impression model, one or more related terms are converted into a set of Kanji characters. In Xu et al. (2006), one or more words that describe the impression of the source word are used as related terms (i.e., impression keywords). Because impression keywords are given manually, users must have a good command of Chinese. In addition, the task of providing impression keywords is expensive. We solve these problems by automatically extracting terms related to the source word from the Web.

Unlike Xu at al. (2006), the language model for the category of the source word is used. For example, if the category is "person", Kanji characters that are often used for personal names in Chinese are preferably used for the transliteration.

Because of the potentially large number of selected candidates, we need to rank the candidates. We model pronunciation, impression, and target language in a probabilistic framework, so that candidates are sorted according to their probability score. In practice, the Kanji characters derived via the impression and language models are used to re-rank the candidates derived via the pronunciation model.

3 Probabilistic Transliteration Model

Given a romanized source word R, a set of related terms W, and the category of the source word C, our purpose is to select the Kanji string K that maximizes $P(K|R, W, C)$, which is evaluated as shown in Equation (1), using Bayes's theorem.

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

Xu et al. (2006) did not consider the category of the source word and computed $P(K|R,W)$.

In the third line of Equation (1), we assume the conditional independence of R, W, and C given K. In the fourth line, we omit $P(R,W,C)$, which is independent of K. This does not affect the relative rank of Kanji strings, when ranked in terms of $P(K|R,W,C)$. If a user intends to select more than one Kanji string, those Ks associated with higher probabilities should be selected. In Figure 1, R, W, and C are "epuson", "喜爱 普及 普通 生动" and " 企業名", respectively, and a K is "埃普松".

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

The implementation of $P(R|K)$ and $P(W|K)$ is the same as in Xu et al. (2006). While $P(R|K)$ has commonly been used in the literature, the basis of $P(W|K)$ should perhaps be explained. $P(W|K)$ is computed using co-occurrence frequencies of each word in W and each character in K, for which we extracted co-occurrences of a word and a Kanji character from a dictionary of Kanji in Chinese. Please see Xu et al. (2006) for details. However, unlike Xu et al. (2006), in which W was provided manually, we automatically extract W from the Web.

While Xu et al. (2006) did not use the language model, we compute $P(C,K)$ by Equation (2).

$$
P(C,K) = P(C) \times P(K|C) \propto P(K|C) \quad (2)
$$

We omit $P(C)$, which is independent of K. Thus, we compute $P(K|C)$, which is the probability that a Kanji string K is selected given category C.

To compute $P(K|C)$, we decompose K into single Kanji characters. We used a character unigram model and produced the following three language models.

- general model: one month of newspaper articles in the PFR corpus[1] were used. In this model, 4 540 character types (12 229 563 tokens) are modeled.
- company model: a list of 22 569 company names in CNLP (Chinese Natural Language Processing)[2] was used. In this model, 2 167 character types (78 432 tokens) are modeled.
- person model: a list of 38 406 personal names in CNLP was used. In this model, 2 318 character types (104 443 tokens) are modeled.

To extract Kanji characters from the above corpus and lists, we performed morphological analysis by SuperMorpho[3] and removed functional words and symbols. While the general model is not adapted to any specific category, the other models are adapted to the company and person categories, respectively. Although the effect of adapting language models has been explored in spoken language processing, no attempt has been made for transliteration.

4 Extracting Related Terms

To extract related terms for a source word, we used Wikipedia[4], which is a free encyclopedia on the Web and includes general words, persons, places, companies, and products, as headwords. We extracted related term candidates for a source word as follows.

1. We consulted the Japanese Wikipedia for the source word and obtained the result page.
2. We deleted HTML tags from the result page and performed morphological analysis by ChaSen[5].
3. We extracted nouns and adjectives as related term candidates.

We used mutual information (Turney, 2001) to measure the degree of relation between the source word and a related term candidate by Equation (3).

$$
I(X,Y) = \log \frac{P(X,Y)}{P(X) \times P(Y)} \quad (3)
$$

[1] http://icl.pky.edu.cn/
[2] http://www.nlp.org.cn/
[3] http://www.omronsoft.com/
[4] http://ja.wikipedia.org/wiki/
[5] http://chasen.naist.jp/hiki/ChaSen/

X and Y denote the source word and a related term candidate, respectively. $P(X)$ and $P(Y)$ denote probabilities of X and Y, respectively. $P(X,Y)$ denotes the joint probability of X and Y.

To estimate the above three probabilities, we followed the method proposed by Turney (2001). We used the Yahoo!JAPAN[6] search engine and replaced $P(A)$ in Equation (3) with the number of pages retrieved by the query A. Here, "A" can be "X", "Y", or "X and Y". Then, we selected up to 10 Ys with the greatest $I(X,Y)$ and translated them into Chinese using the Yahoo!JAPAN machine translation system.

Table 1 shows examples of related terms for the source word "ミサ (mass)", such as "典礼 (ceremony)" and " 奉献 (dedication)". Irrelevant candidates, such as " 会 (meeting)" and "こと (thing)", were discarded successfully.

Table 1: Example of related terms for "ミサ (mass)".

Extracted related terms		Discarded candidates	
Japanese	English	Japanese	English
典礼	ceremony	会	meeting
奉献	dedication	こと	thing
司教	bishop	会議	meeting
教会	church	参加	join

5 Experiments

5.1 Method

To evaluate the effectiveness of the related term extraction in the transliteration, we compared the accuracy of the following three methods.

- A combination of the pronunciation and language models that does not use the impression model, $P(W|K)$, in Equation (1),
- Our method, which uses Equation (1) and uses automatically extracted related terms as W,
- Equation (1), in which manually provided impression keywords are used as W.

To make the difference between the second and third methods clear, we use the terms "related term (RT)" and "impression keyword (IK)" to refer to words provided automatically and manually, respectively. Then, we call the above three methods "PL", "PL+RT", and "PL+IK", respectively. PL and PL+IK are the lower bound and the upper bound of the expected accuracy, respectively. PL+IK is the same as in Xu et al. (2006), but the language model is adapted to the category of source words.

To produce test words for the transliteration, we first collected 210 Katakana words from a Japanese–Chinese dictionary. These 210 words were also used by Xu et al. (2006) for experiments. We then consulted Wikipedia for each of the 210 words and selected 128 words that were headwords in Wikipedia, as test words. Details of the 128 test words are shown in Table 2.

Table 2: Categories of test words.

Category	#Words	Example word		
		Japanese	Chinese	English
General	24	エンジェル	安琪儿	angel
Company	35	インテル	英特尔	Intel
Product	27	アウディ	奥迪	Audi
Person	13	ショパン	肖邦	Chopin
Place	29	オハイオ	俄亥俄	Ohio

We selectively used the three language models explained in Section 3. We used the general model for general words. We used the company model for company and product names, and used the person model for person and place names. A preliminary study showed that the language model adaptation was generally effective for transliteration. However, because the focus of this paper is the related term extraction, we do not describe the evaluation of the language model adaptation.

Two Chinese graduate students who had a good command of Japanese served as assessors and produced reference data, which consisted of impression keywords used for PL+IK and correct answers for the transliteration. Neither of the assessors was an author of this paper. The assessors performed the same task for the 128 test words independently, to enhance the objectivity of the evaluation.

We produced the reference data via the following procedure that is the same as that of Xu et al. (2006).

First, for each test word, each assessor provided one or more impression keywords in Chinese. We did not restrict the number of impression key-

[6] http://www.yahoo.co.jp/

words per test word; the number was determined by each assessor. We provided the assessors with the descriptions for the test words from the source Japanese–Chinese dictionary, so that the assessors could understand the meaning of each test word.

Second, for each test word, we applied the three methods (PL, PL+RT, and PL+IK) independently, which produced three lists of ranked candidates.

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 three ranked lists. We merged these candidates, removed duplications, and sorted the remaining candidates by character code. The assessors judged the correctness of up to 300 candidates for each test word. The average number of candidates was 36 976.

The resultant reference data were used to evaluate the accuracy of each 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.

For each test word, there was more than one type of "correct answer", as follows:

(a) transliteration candidates judged as correct by either of the assessors independently,

(b) transliteration candidates judged as correct by both assessors,

(c) transliteration defined in the source Japanese–Chinese 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. In (b), where the assessors did not disagree about the correctness, the coverage of the correctness and the objectivity are in between.

The number of test words was 128 for both (a) and (c), but 76 for (b). The average numbers of correct answers were 1.65, 1.04, and 1 for (a), (b), and (c), respectively.

5.2 Results and Analyses

Table 3 shows the average rank of correct answers for different cases. Looking at Table 3, for certain categories, such as "Place", when the impression model was used, the average rank was low. However, on average, the average rank for PL+RT was lower than that for PL+IK, but was higher than that for PL, irrespective of the answer type.

Figures 2 and 3 show the distribution of correct answers for different ranges of ranks, using answer types (a) and (c) in Table 3, respectively. Because the results for types (a) and (b) were similar, we show only the results of type (a), for the sake of conciseness. In Figure 2, the number of correct answers in the top 10 for PL+RT was smaller than that for PL+IK, but was greater than that for PL.

In Figure 3, the number of correct answers in the top 10 for PL+RT was greater than those for PL and PL+IK. Because in Figure 3, the correct answers were defined in the dictionary and were independent of the assessor judgments, PL+IK was not as effective as in Figure 2.

In summary, the use of automatically extracted related terms was more effective than the method that does not use the impression model. We also reduced the manual cost of providing impression keywords, while maintaining the transliteration accuracy.

Table 4 shows examples of related terms or impression keywords for answer type (c). In Table 4, the column "Rank" denotes the average rank of correct answers for PL+RT and PL+IK, respectively. For "ミサ (mass)", the rank for PL+RT was higher than that for PL+IK. However, for "カタール (the State of Qatar)", the rank for PL+RT was lower than that for PL+IK. One reason for this is that most related terms for PL+RT were names of countries that border Qatar, which do not describe Qatar well, compared with impression keywords for PL+IK, such as "沙漠 (desert)" and " 石油 (oil)". This example indicates room for improvement in the related term extraction algorithm.

6 Conclusion

For transliterating foreign words into Chinese, the pronunciation of a source word is spelled out with Kanji characters. Because Kanji is an ideogrammatic script, different Kanji characters are associ-

Table 3: Average rank of correct answers for different methods in different cases.

Category	Answer type (a)			Answer type (b)			Answer type (c)		
	PL	PL+RT	PL+IK	PL	PL+RT	PL+IK	PL	PL+RT	PL+IK
General	189	165	167	44	49	52	84	61	65
Company	232	208	203	33	29	27	317	391	325
Product	197	175	166	34	27	21	313	198	198
Person	98	69	44	4	4	4	114	154	75
Place	85	133	95	13	14	16	76	98	89
Avg.	160	150	135	26	25	24	181	160	150

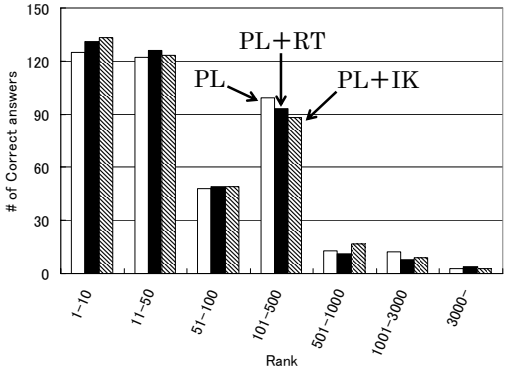

Figure 2: Rank for correct answer type (a).

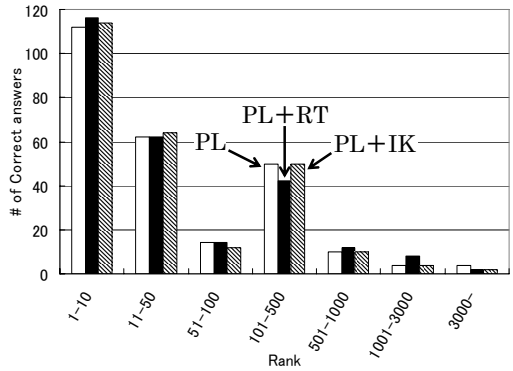

Figure 3: Rank for correct answer type (c).

Table 4: Examples of related terms and impression keywords used for experiments.

Source word	Answer	Method	Rank	Examples of related terms or impression keywords
ミサ (mass)	弥撒	PL+RT	8	典礼 (ceremony), 主教 (bishop), 奉献 (dedication), 教会 (church)
		PL+IK	10	典礼 (ceremony), 主教 (bishop), 信仰 (belief), 教会 (church)
カタール (State of Qatar)	卡塔尔	PL+RT	103	科威特 (State of Kuwait), 也门 (Republic of Yemen)
		PL+IK	61	阿拉伯 (Arab), 沙漠 (desert), 石油 (oil), 干燥 (dryness)

ated with the same pronunciation, but can potentially convey different meanings and impressions. In this paper, to select appropriate characters for transliterating into Chinese, we automatically extracted related terms for source words using the Web. We showed the effectiveness of our method experimentally.

References

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.

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

Peter D. Turney. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning*, pages 419–502.

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.

LiLi Xu, Atsushi Fujii, and Tetsuya Ishikawa. 2006. Modeling Impression in Probabilistic Transliteration into Chinese. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 242–249.