An Approach to Build a Proper Noun Dictionary for Record Linkage across Humanities Databases in Different Languages

Yuting SONG† Taisuke KIMURA† Biligsaikhan BATJARGAL‡ Akira MAEDA‡†

† Graduate School of Information Science and Engineering, Ritsumeikan University

1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577 Japan

‡ Research Organization of Science and Engineering, Ritsumeikan University

1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577 Japan

‡† College of Information Science and Engineering, Ritsumeikan University

1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577 Japan

E-mail: †{gr0260ff, is0013hh}@ ed.ritsumei.ac.jp, ‡biligee@fc.ritsumei.ac.jp, ‡†amaeda@is.ritsumei.ac.jp

Abstract This paper proposes a method to build a proper noun dictionary that contains proper noun and its transliterated word pairs, which can be used for record linkage across humanities databases in different languages. This method is based on an observation that the corresponding word of a proper noun in a target language is almost the same as transliteration of that proper noun. In this paper, the proposed method is demonstrated on Ukiyo-e databases in Japanese (source language) and English (target language). We extract transliterated words from the titles of Ukiyo-e prints in English and back-transliterate them into their original Japanese words by using a kanji dictionary and Ukiyo-e related dictionaries and encyclopedias. Finally, we evaluate how the newly built proper noun dictionary affects the accuracy of record linkage between Ukiyo-e databases in Japanese and English.

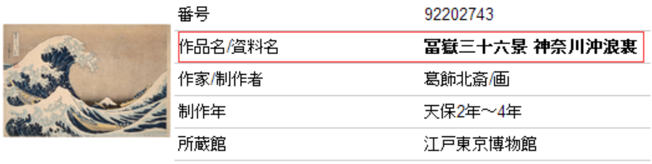
Keyword Cross-language record linkage, Digital library, Japanese art, Humanities databases

1. Introduction

Cross-Language Record Linkage is a task of finding the pairs of records that refer to the same entity across multiple databases in different languages. In this task, the metadata elements of a record in original language need to be translated into the target language in order to match records in that language. Proper nouns in the metadata elements, such as personal names, place names, are more representative features than other words. However, it is difficult to obtain proper nouns and their corresponding transliterations by dictionary lookup because they are often domain specific and obsoleted, especially in humanities areas.

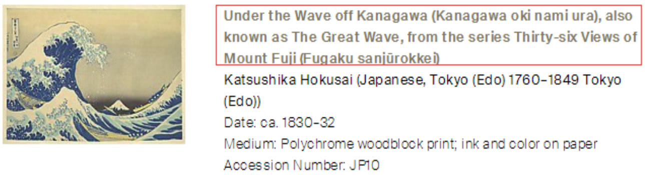
This paper proposes a method to build a proper noun dictionary that contains proper noun and its transliterated word pairs, which can be used for record linkage [1][2][3] across humanities databases in different languages. This method is based on an observation that the corresponding word of a proper noun in a target language is almost the same as transliteration of that proper noun.

In this paper, we use Ukiyo-e databases in Japanese and English to demonstrate the proposed method. The Ukiyo-e, Japanese traditional woodblock printing, is known as one of the popular arts of the Edo period (1603-1868). Many libraries, museums and organizations in Japan as well as in other countries hold numerous Ukiyo-e prints. These prints have been digitized and exhibited on the Internet with textual metadata in different languages [4]. As shown in the red boxes of Figure 1 and Figure 2, the same Ukiyo-e prints have different titles, e.g. Japanese title or English translated title. Thus, it is suitable to demonstrate our proposed method on Ukiyo-e databases in Japanese and English. Besides, we use the titles of Ukiyo-e prints for the task of identifying the same prints between the databases in Japanese and English because the title is a more definite metadata element than others, such as year of creation, published place, etc. In this task, the titles of Ukiyo-e prints in Japanese (source language) need to be translated into English (target language) to match the prints in English.



番号	92202743
作品名/資料名	**冨嶽三十六景 神奈川沖浪裏**
作家/制作者	葛飾北斎/画
制作年	天保2年～4年
所蔵館	江戸東京博物館

Figure 1. An example of the Edo-Tokyo Museum



Under the Wave off Kanagawa (Kanagawa oki nami ura), also known as The Great Wave, from the series Thirty-six Views of Mount Fuji (Fugaku sanjūrokkei)

Katsushika Hokusai (Japanese, Tokyo (Edo) 1760–1849 Tokyo (Edo))

Date: ca. 1830–32

Medium: Polychrome woodblock print; ink and color on paper

Accession Number: JP10

Figure 2. An example of the Metropolitan Museum of Art

The remainder of the paper is organized as follows. Section 2 introduces the related work for acquisition of original word and its transliterated word pairs in several languages. The proposed method is described in Section 3. Experimental setup and evaluation of the performance are presented in Section 4. Finally, the conclusion follows in Section 5.

2. Related Work

In the past, much research has been conducted for acquisition of original word and its transliterated word pairs in many languages, such as English and Chinese [5] [6], English and Korean [7] [8], as well as English and Japanese [9]. In most approaches, they focus on the transliteration and the back-transliteration between English words and foreign words (e.g. English loanwords in Japanese), in which English is the source language and non-English languages are the targets.

However, the proposed method in this paper aims to obtain non-English word and its corresponding English transliterated word pairs, in which non-English language is the source and English is the target language. This task is more challenging than the back-transliteration between English words and foreign words, since each character of foreign words can be corresponded to one or more English letters of corresponding English words. However, in the task of the back-transliteration between non-English words and English transliterated words, several letters in English transliterated word can be represented as a single non-English character. Some examples are shown in Figure 3.

The example on the left side of Figure 3 illustrates the back-transliteration from loanword “クリーム” in Japanese to English word “cream”. In this process, each katakana character corresponds to one or more English letters. For example, “ク” corresponds to “c” in English word “cream”. However, the right side illustrates the back-transliteration from English transliterated word “ono” to its corresponding Japanese words. In this example, “o” and “no” have their corresponding Japanese kanji “小” and “野” (“小野” is a place name). Besides, “o” and “no” can share one corresponding Japanese kanji “斧” (an axe).

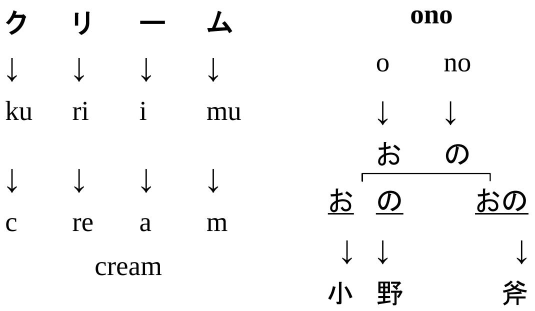


Figure 3. Examples of the back-transliteration between an English word and a Japanese foreign word and the back-transliteration between a Japanese word and English transliterated word

3. Proposed method

In the proposed method, we identify proper nouns by hypothesizing that the corresponding words of transliterated words in the target language are proper nouns. Based on this idea, firstly, we extract transliterated words from the English titles of Ukiyo-e prints. In this step, we identify transliterated words by a language dictionary lookup since transliterated words are not always appeared in a language dictionary, especially in humanities areas where proper nouns are domain-specific or obsoleted. Then, the next step is to convert these transliterated words into their corresponding words in Japanese. This back-transliteration process is to recover the original words from transliterated words. After that, candidate proper noun and its transliterated word pairs are post-processed in order to build a more accurate proper noun dictionary. The process of back-transliteration and post-processing will be described in details in the next subsection. The overall process is summarized in Figure 4.

3.1 Back-transliteration

For building a proper noun dictionary that contains proper noun and its transliterated word pairs, transliterated words in English are needed to be converted to their original words in Japanese. This process is called back-transliteration that is the reverse process of transliteration. For example, transliterated word “sumida” in English is needed to be converted back to “隅田” (the name of a river) in Japanese.

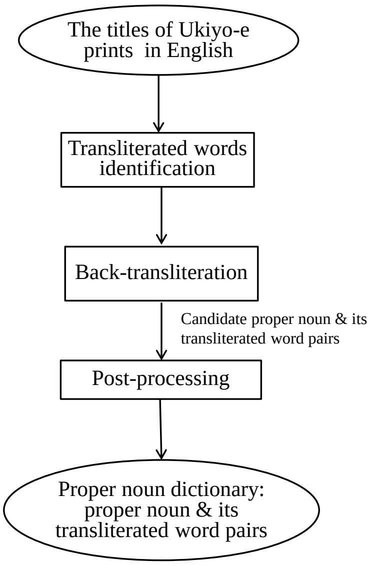

The titles of Ukiyo-e prints in English
Transliterated words identification
Back-transliteration
Candidate proper noun & its transliterated word pairs
Post-processing
Proper noun dictionary: proper noun & its transliterated word pairs

Figure 4. The overall process of building a proper noun dictionary

Figure 5 shows the back-transliteration process in detail. Transliterated words in English titles are converted to corresponding Japanese hiragana sequences based on Hepburn Romanization[1] system. In order to find the corresponding Japanese words for hiragana sequences, an aligned Japanese word and its hiragana sequence pairs are needed. Thus, we align Japanese word and its hiragana sequence pairs of the entries in Ukiyo-e related dictionaries and encyclopedias character by character. In this character alignment process, the pronunciation of each character in Japanese words is determined by a Japanese kanji pronunciation dictionary lookup. According to this alignment results, the Japanese hiragana sequences of transliterated words are used to match or match partially the hiragana sequences of the entries in Ukiyo-e related dictionaries and encyclopedias for obtaining their corresponding Japanese words. Finally, candidate Japanese word (proper noun) and its transliterated word pairs are generated through the hiragana sequences.

The back-transliteration process for some transliterated words in the English titles of Ukiyo-e prints is illustrated in Figure 6. In this example, firstly, a transliterated word “Sumida” is converted into its corresponding Japanese hiragana sequence “すみだ”. According to the character alignment result “隅｜田｜川 -- すみ｜だ｜がわ”[2] in Ukiyo-e related dictionaries and encyclopedias, “すみだ” is used to match partially the hiragana sequences “すみだがわ” of the entries “隅田川－すみだがわ” for obtaining its corresponding Japanese word “隅田”.

3.2 Post-processing

In the back-transliteration process, most transliterated words can be converted into one or more original Japanese words that might be proper nouns. Among 247 transliterated words in our experimental result (the experiment will be described in details in Section 4), 138 words (55.9%) have more than one corresponding Japanese words. However, some of them are ordinary Japanese words which are not proper nouns. For example, a transliterated word “fudō” can be back-transliterated as “不動” (motionless), “歩堂” or “武道” (martial arts), where the two words are not proper noun. Thus, it is vital to select actual proper nouns among the candidate proper nouns in order to improve the quality of the proper noun dictionary. For this task, the proper nouns are determined by Japanese-English bilingual dictionary lookup by assuming any word that is not listed in the bilingual dictionary is a proper noun, which is depicted as post-processing based on dictionary lookup in Figure 4.

Besides, some transliterated words might be back-transliterated into incorrect Japanese words. For example, transliterated word “yoshino” can be back-transliterated into “吉野” (a place name), “葭の” and “吉の”. Among them, “葭の” and “吉の” are not correct Japanese words. Such an error can be partly eliminated by adding some constraints on word composition. We calculate the Japanese kanji and kana distributions of proper nouns composition from a Japanese proper noun list in humanities domains. Majority of Japanese proper nouns consist of kanji. A small number of proper nouns contain Japanese kana in the middle of a proper noun, such as “七里ヶ浜” (a beach called Shichirigahama) and “程ヶ谷” (a valley called Hodogaya). Therefore, incorrect Japanese proper nouns could be identified by the position of kana. If a Japanese kana appears at the end of a word, that word might be a non-proper noun. For example, “葭の” and “吉の” might be non-proper nouns because the Japanese kana “の” appears at the end of the words.

[1] https://en.wikipedia.org/wiki/Hepburn_romanization

[2] “隅田川” is a river called Sumida; “すみだがわ” is the Japanese pronunciation of “隅田川”.

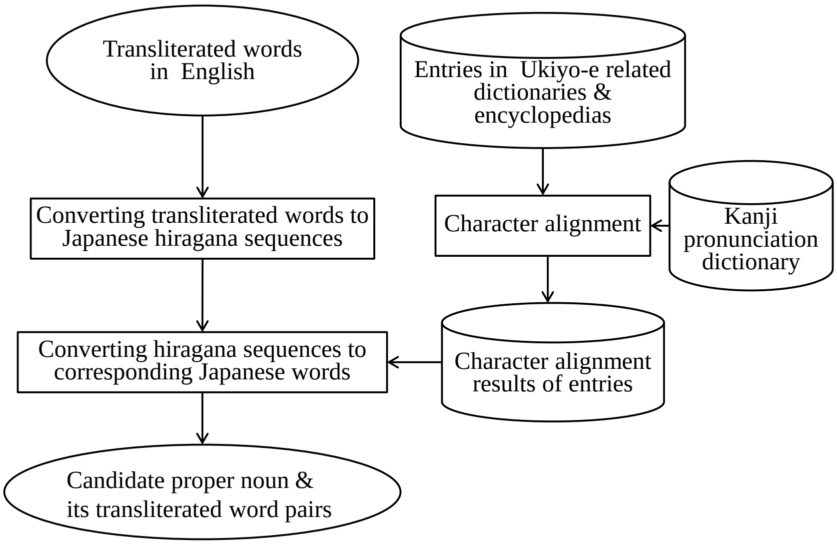

Transliterated words in English
Entries in Ukiyo-e related dictionaries & encyclopedias
Converting transliterated words to Japanese hiragana sequences
Character alignment
Kanji pronunciation dictionary
Converting hiragana sequences to corresponding Japanese words
Character alignment results of entries
Candidate proper noun & its transliterated word pairs

Figure 5. The back-transliteration process

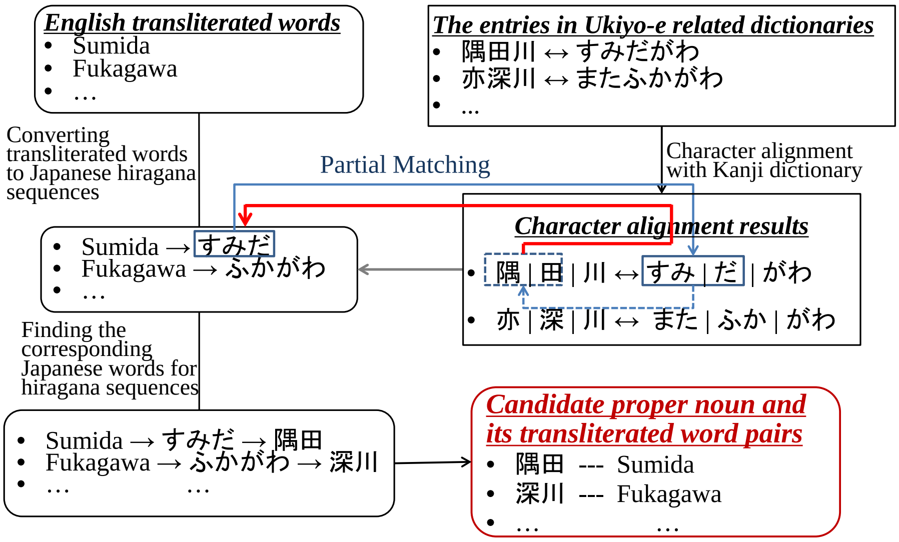

English transliterated words
• Sumida
• Fukagawa
• …
The entries in Ukiyo-e related dictionaries
• 隅田川 ↔ すみだがわ
• 亦深川 ↔ またふかがわ
• ...
Converting transliterated words to Japanese hiragana sequences
Partial Matching
Character alignment with Kanji dictionary
Character alignment results
• 隅 | 田 | 川 ↔ すみ | だ | がわ
• 亦 | 深 | 川 ↔ また | ふか | がわ
• Sumida → すみだ
• Fukagawa → ふかがわ
• …
Finding the corresponding Japanese words for hiragana sequences
• Sumida → すみだ → 隅田
• Fukagawa → ふかがわ → 深川
• … …
Candidate proper noun and its transliterated word pairs
• 隅田 --- Sumida
• 深川 --- Fukagawa
• … …

Figure 6. The back-transliteration process of some transliterated words in the titles of Ukiyo-e prints in English

4. Experiments

In this section, we discuss the experimental setup and evaluation for the proposed method.

4.1 Experimental setup

Two preliminary experiments are conducted to evaluate the proposed method. The first one is to build a proper noun dictionary. The second one is to use the proper noun dictionary that is obtained from the first experiment for the task of record linkage across humanities databases in different languages.

The experimental data consists of 45 Japanese titles of the artist Katsushika Hokusai's Ukiyo-e prints in the Edo-Tokyo Museum [3] and 437 English titles in the Metropolitan Museum of Art [4], in which each Japanese title has at least one corresponding English title.

In the process of building a proper noun dictionary, transliterated words are identified from 437 English titles of Ukiyo-e prints by using Merriam-Webster's Learner's Dictionary [5]. Besides, the Japanese word and its hiragana sequence pairs of the entries in Ukiyo-e related dictionaries and encyclopedias are aligned by using the Kanji dictionary in Kanjijite.net [6]. Ukiyo-e related

[3] http://digitalmuseum.rekibun.or.jp/index.html

[4] http://www.metmuseum.org/

[5] http://www.dictionaryapi.com/products/api-learners-dictionary.htm

[6] http://www.kanjijiten.net/index.html

dictionaries and encyclopedias include 《浮世絵百科》, 《浮世絵大事典》, 《浮世絵事典》[7], 《浮世絵鑑賞基礎知識》[8], 《浮世絵百科（画題）》[9] and 《浮世絵大武者絵展》[10]. Finally, the Japanese proper nouns are selected from candidate proper noun and its transliterated word pairs by using New College Japanese-English Dictionary[11] in the post-processing process.

4.2 Evaluation

In the first experiment, 890 candidate pairs of proper noun and its transliterated word are obtained after the back-transliteration process. We check these candidates manually and find that there are 571 correct proper nouns among them. In the post-processing step, 720 words are determined as proper nouns from the 890 candidates. Among these determined proper nouns, the number of correctly determined proper nouns is 559. The experimental result is evaluated based on precision and recall rates:

$$\text{Precision} = \frac{\textit{the number of correctly determined proper nouns}}{\textit{the number of determined proper nouns}} \quad (1)$$

$$\text{Recall} = \frac{\textit{the number of correctly determined proper nouns}}{\textit{the number of correct proper nouns}} \quad (2)$$

As shown in the first row of Table 1, the proposed method achieves 72.4% of precision and 97.9% of recall. The precision rate is improved by adding the constraint on the appearance of kana at end of Japanese word, as shown in the second row of Table 1.

Table 1. The experimental results of building a proper noun dictionary

Experimental method	Precision	Recall
Baseline	559 / 720 = 72.4%	559 / 571 = 97.9%
Baseline + Constraint	559 / 685 = 81.6%	559 / 571 = 97.9%

However, a small number of proper nouns are incorrectly determined as non-proper nouns because they are frequently used proper nouns that are listed in bilingual dictionary, such as "江戸" (old name of Tokyo), "日本" (Japanese), etc.

In the second experiment, the proper noun dictionary that is obtained from the first experiment is used to identify the same Ukiyo-e prints between the databases in Japanese and English [10]. We have successfully identified the 42 Ukiyo-e prints out of 45, as shown in row 2 of Table 2. The experimental result is evaluated based on the accuracy in top 1 and top 5, which means the correct Ukiyo-e prints rank in the top 1 or within top 5 of the identified results. Compared with the result of our previous approach [10] that are showed in row 1 of Table 2, both the accuracy in top 1 and top 5 are improved by utilizing the proper noun dictionary that is built by the proposed method.

Table 2. The experimental results of identifying the same Ukiyo-e prints between the databases in Japanese and English

Experimental method	Accuracy in top 1	Accuracy within top 5
Our previous approach [10]	35 / 45 = 77.8%	39 / 45 = 86.7%
Utilizing the proper noun dictionary	42 / 45 = 93.3%	42 / 45 = 93.3%

The experimental results show that, the proposed method to build a proper noun dictionary can improve the identification performance of the same Ukiyo-e prints of our previous approach [10] by determining and transliterating the proper nouns in the titles more precisely. For example, the same Ukiyo-e prints with the Japanese title "江都駿河町三井見世略図" and English title "Mitsui Shop at Surugachô in Edo" are identified successfully, since the proper nouns "江都" (the name of an ancient Japanese city), "駿河町" (a place name) and "三井" (a shop name) is determined and transliterated correctly by the proposed method, which was failed in our previous approach [10]. A few Ukiyo-e prints are incorrectly identified because some non-proper nouns are determined as proper nouns in the preceding task. For example, a non-proper noun "諸人" (many people) in the Japanese title "諸人登山" (groups of mountain climbers) is incorrectly determined as a proper noun.

5. Conclusion

In this paper, we proposed a method to build a proper noun dictionary that contains proper noun and its transliterated word pairs. The proposed method can be used to deal with the task of proper noun transliteration for record linkage across humanities databases in different languages. In the future, we plan to extend the proposed method to other humanities databases.

[7] The encyclopedias 《浮世絵百科》, 《浮世絵大事典》, 《浮世絵事典》 contains the titles, artists, subjects, publishers and some other information of Ukiyo-e prints.

[8] 《浮世絵鑑賞基礎知識》 is a dictionary of the basic knowledge for appreciating Ukiyo-e prints.

[9] 《浮世絵百科（画題）》 is an encyclopedia of the titles of Ukiyo-e prints.

[10] 《浮世絵大武者絵展》 is an encyclopedia of warriors in Ukiyo-e prints.

[11] http://ejje.weblio.jp/category/dictionary/kenje

References

[1] I. P. Fellegi and A. B. Sunter, “A Theory for Record Linkage”, Journal of the American Statistical Association. Vol. 64, No.328, pp. 1183-1210, 1969.

[2] S. Sarawagi and A. Bhamidipaty, “Interactive Deduplication using Active Learning”, Proc. of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 269-278, 2002.

[3] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar and S. Fienberg, “Adaptive Name Matching in Information Integration”, IEEE Intelligent Systems. Vol.18, No.5, pp. 16-23, 2003.

[4] B. Batjargal, T. Kuyama, F. Kimura, and A. Maeda, “Identifying the Same Records across Multiple Ukiyo-e Image Databases Using Textual Data in Different Languages", Proc. of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014), pp. 193-196, 2014.

[5] H. Chen, S. Huang, Y. Ding, and S. Tsai. 1998. “Proper Name Translation in Cross-language Information Retrieval”, Proc. of 17th COLING and 36th ACL, pp. 232-236, 1998.

[6] C. Lee and J. S. Chang, “Acquisition of English-Chinese Transliterated Word Pairs from Parallel-Aligned Texts”, Proc. of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond - Volume 3, pp. 96-103, 2003.

[7] K.S. Jeong, S.H. Mayeng, J.S. Lee and K.S. Choi, “Automatic identification and back-transliteration of foreign words for information retrieval”, Information Processing & Management, Vol.35, No.4, pp. 523-540, 1999.

[8] B. Kang and K. Choi. “English-Korean Automatic Transliteration/Back-transliteration System and Character Alignment”. Proc. ACL, pp. 17–18, 2000.

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

[10] T. Kimura, B. Batjargal, F. Kimura and A. Maeda, “Finding the Same Artworks from Multiple Databases in Different Languages,” In Conference Abstracts of Digital Humanities 2015, 2015.