

Chapter 3

Linking Ukiyo-e Records across Languages:
An Application of Cross-Language Record
Linkage Techniques to Digital Cultural Collections

Yuting SONG

The work described in this chapter was jointly done by Dr. Biligsaikhan Batjargal and Professor Akira Maeda at the College of Information Science and Engineering, Ritsumeikan University.

Ukiyo-e is well-known as a traditional Japanese art form and is one of the popular styles of the Edo period (1603–1868). Not only in Japan but also in many Western countries, there are many museums, libraries, and galleries that have digitized Ukiyo-e block prints. In different countries, digital cultural collections are described using metadata in different languages, and for example, digitized Ukiyo-e prints are found on the internet with metadata in Japanese, English, and Dutch.







| | Metadata | | Language |
|--|--|------------------------|--------------------------------|
| | Title | Artist | |
| 江戸東京博物館 (Japan) |  雪月花 淀川 | 葛飾北斎 | Identical ukiyo-e prints |
| |  凱風快晴 | 葛飾北斎 | |
| Metropolitan Museum of Art (United States) |  Moonlight on the Yodo River, from the series Snow, Moon, and Flowers | Katsushika Hokusai | English |
| |  Morning Mist at Mishima | Utagawa Hiroshige (I) | Identical ukiyo-e prints |
| Rijksmuseum (Netherlands) |  Helder weer en een zuidelijke wind | Katsushika Hokusai | |
| |  Mishima in ochtendmist | Hiroshige (I), Utagawa | |

Figure 1. Digital versions with metadata in different languages
Source: Author

We can see that identical Ukiyo-e prints could be described in different languages like the examples in Figure 1. The metadata records refer to the same Ukiyo-e prints, but the metadata is in Japanese, English, and Dutch. The purpose of our work is to find the identical Ukiyo-e prints by using the textual metadata in different languages, for example, on the titles and artist names in Japanese and English as in Figure 2.



| 作品名 | 作者 | | Title | Artist |
|------------------|-----------|---|---|--------------------|
| 富嶽三十六景 神奈川沖浪裏 | 葛飾北斎 |  | Under the Wave off Kanagawa, from the series Thirty-six Views of Mount Fuji | Katsushika Hokusai |
| 富嶽三十六景 深川万年橋下 | 葛飾北斎 | | Snow on the Sumida River, from the series, Snow, Moon, and Flowers | Katsushika Hokusai |
| 日本橋 朝之景 | 歌川広重 (初代) |  | Morning View of Nihonbashi | Utagawa Hiroshige |
| 雪月花 隅田 | 葛飾北斎 | | | |

Figure 2. The examples of the identical Ukiyo-e metadata records from different digital collections in Japanese and English

Source: Author

This research could help people to search the Ukiyo-e prints, regardless of language and it could also help metadata creators enrich metadata in other languages.

Cross-language Record Linkage

Cross-language record linkage is used to identify record pairs that refer to the same real-world entities across data sources in different languages. Given the metadata records in different languages, the key step is to measure the similarities between metadata in the different languages, which is the across-language metadata similarity calculation. Then, based on these metadata similarities, the record pairs are classified

into matches or non-matches by using a decision model.

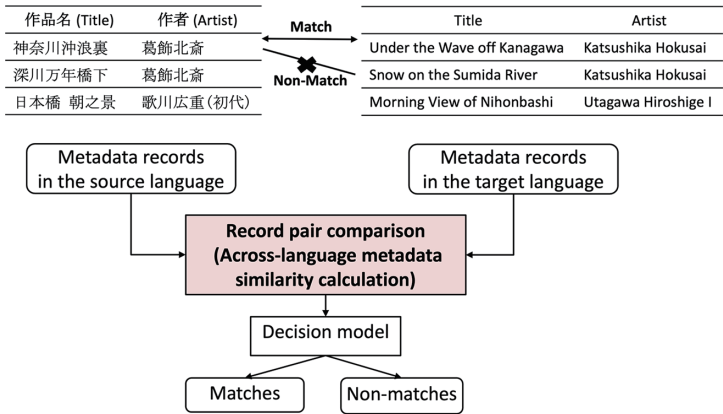


Figure 3. Calculating metadata similarities in different languages
Source: Author

The challenge of cross-language record linkage is how to calculate the metadata similarity in different languages. As shown in Figure 3, we deal with this problem from two directions. One is a translation-based method, which uses machine translation to overcome the language barriers, and the other is the method without translation. The metadata are translated from one language to the other using the translation-based method. As shown in Figure 4, the metadata are translated from Japanese to English by means of machine translation and then compared within the same language. For the second method, without using translation, we use bilingual word embeddings to represent words in metadata within one vector space, and then calculate metadata similarity.

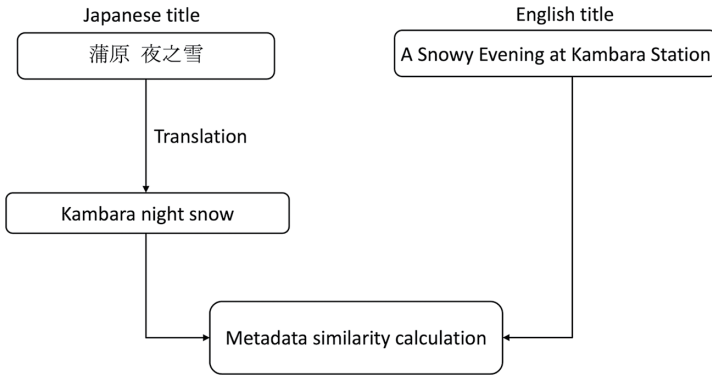


Figure 4. Metadata similarity calculation using translation-based method

Source: Author

Let me briefly explain word embedding and bilingual word embedding. Embeddings are the distributed vector representations of words, which are dense, low dimensional, and real value vectors. Usually, the dimension of the word vector is fixed, for example, 200 or 300. Figure 5 is an example of the word vector to represent the English word “storm.”

- **The dimension of the vector is hundreds (eg. 200, 300)**
- **e.g. $vector(storm) = [0.23, 0.44, -0.76, 0.33, 0.19, \dots]$ dim = 200**

Figure 5. Vector for the word “storm”

Source: Author

One of the advantages of word embeddings is that semantically similar words are close in the vector space, and the semantic similarity between words can be calculated using this word vector.

In bilingual word embedding, the cross-lingual vector representations of words can be obtained by linear mapping between monolingual word

embeddings.

Figure 6 shows a cross-lingual word embedding space. In this space, words such as the Japanese word “*sekai*” and its corresponding English word “world” have a closer distance represented by the close vectors.

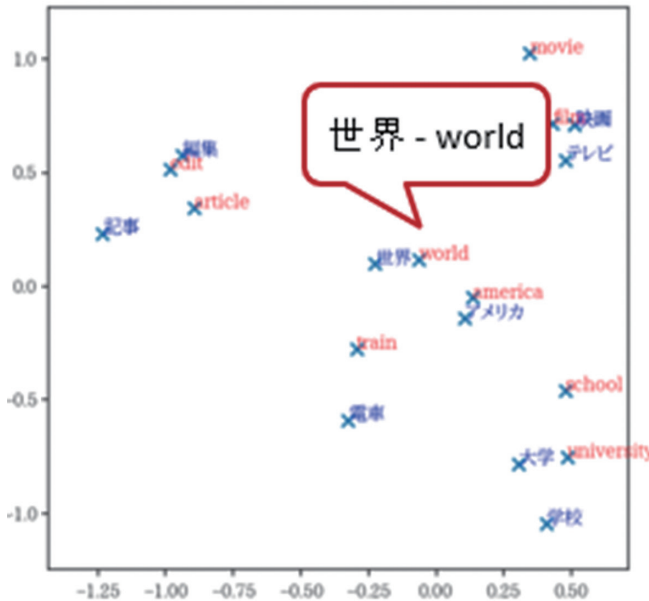


Figure 6. Words in different languages that are similar have close vectors.
Source: Author

Because bilingual word embeddings can capture the semantic meaning of words across languages, we use these vectors to represent the words in metadata, and then calculate the metadata similarities in different languages.

The process for calculating the similarity of Ukiyo-e titles using bilingual word embeddings involves the following steps. Firstly, we

will represent all the words in the source and target languages using bilingual word embedding. For each word in the title in the source language, we calculate the cosine similarity with each word in the title in the target language, as shown in the equation at the bottom of Figure 7. At the same time, we also use the romanization of this word to calculate the cosine similarity with each word in the title in the target language. Accordingly, we use the maximum similarity score as the contribution of this word to the title similarity.

- For each word w_i^{ts} in the title in the source language
 - It is used to calculate the cosine similarity with each word w_j^{tr} in the title in the target language
 - The romanization of word w_i^{ts} is also used to calculate the cosine similarity with each word w_j^{tr} in the title in the target language
 - The maximum similarity score is used as the contribution of w_i^{ts} to the title similarity

$$S_{W2W_Mat}(t_s, t_T) = \sum_i^{N_{t_s}} \max_{w_j^{tr} \in t_T} [\cosine(w_i^{ts}, w_j^{tr}), \cosine(w_{ri}^{ts}, w_j^{tr})]$$

Figure 7. Metadata similarity calculation

Source: Author

Finally, the title similarity is calculated by using this formula and the sum of all the words that contribute to the title similarity. The example in Figure 8 shows the process of how our method calculates Ukiyo-e title similarity in Japanese and English. The arrows represent the similarity calculation between two-word vectors. Each word is used to calculate the cosine similarity with the word in the English title. The words in English titles are also represented by the bilingual word embedding. The red arrows represent the maximum similarities of Japanese words, which are the contributions of Japanese words to the title similarity calculation.

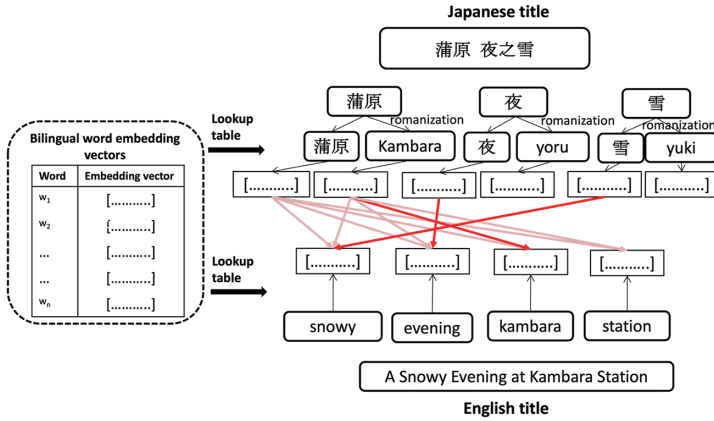


Figure 8. Example of title similarity calculation using bilingual word embeddings
Source: Author

Next, we did some experiments to find identical Ukiyo-e prints between databases in Japanese and English. We used the title of Ukiyo-e prints to evaluate our methods. To reduce the number of record pairs to be compared, we filtered the candidate record pairs by using the artists' names of Ukiyo-e prints. We collected 203 Japanese records from the Edo Tokyo Museum and 3,398 English records from the Metropolitan Museum of Art. Each Japanese Ukiyo-e metadata record has a corresponding English Ukiyo-e metadata record in this English data set, which means they referred to the identical prints that we wanted to find.

For the machine translation-based method, we translated Ukiyo-e titles from Japanese to English by using three well-known online machine translation systems: Microsoft Translator, Google Translate, and DeepL Translator. For the method without translation, we need to learn bilingual word embedding, which requires monolingual word

embeddings and bilingual word pairs. To learn the monolingual word embeddings, we use Japanese and English Wikipedia articles and the Word2Vec toolkit. For the bilingual word pairs, we use 9000 common Japanese words and their English translations to learn the mapping between the Japanese word embedding space and English word embedding space. We followed the training setup in this paper [1].

The results for finding identical Ukiyo-e records are shown in Table 1.

| Methods | MAP | Top-1 precision | Top-3 precision |
|---------------|--------|-----------------|-----------------|
| MS | 0.4949 | 0.3839 | 0.5663 |
| Google | 0.5750 | 0.4478 | 0.6522 |
| DeepL | 0.5114 | 0.3922 | 0.5862 |
| BiWE-w2w | 0.2698 | 0.2007 | 0.3002 |
| BiWE-w2w+roma | 0.5338 | 0.4660 | 0.5721 |

Table1. Experimental results

Among all the results, we can see that Google Translator obtained the best performance. Within the translation-based methods, we can see that the result is actually influenced by the different translation systems.

If we compare the method based on translation and the method without translation, although the result of the bilingual word embedding-based method is worse than that of the translation-based method, it is a promising result because it can be used for other low-resource language pairs where machine translation is not available.

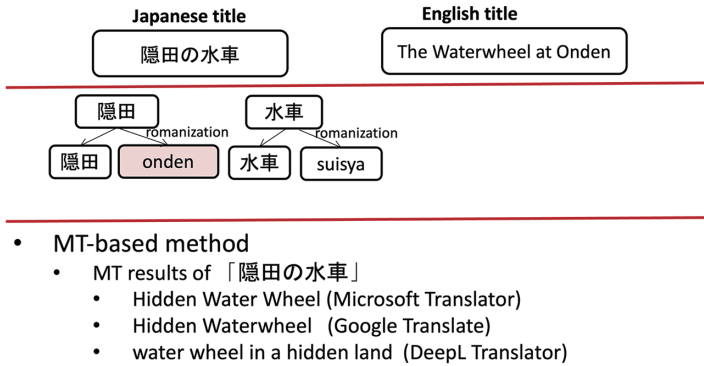


Figure 9. An example demonstrating that the bilingual word embedding-based method is better than the MT-based method

Source: Author

Figure 9 shows how using the romanization of Japanese words in the bilingual word embedding-based method contributes to the title similarity calculation. The corresponding Japanese word in this English title is romanized as “Oden,” but all the machine translation systems translated this word as “hidden” or “hidden land.” Thus, in this case, the machine translation result could not match its corresponding word in the English title.

We also would like to discuss the limitations of our approach. Because our method relies on the text for metadata, our methods failed in cases when the corresponding English title is an inadequate translation. For instance, as illustrated in Figure 10, these two identical Ukiyo-e prints’ the English title did not contain the first two Japanese words in the Japanese title. Our method failed in cases like this.

Chapter 3

Linking Ukiyo-e Records across Languages: An Application of Cross-Language Record Linkage Techniques to Digital Cultural Collections



Figure 10. An example of the limitation of our approach

Source: Author

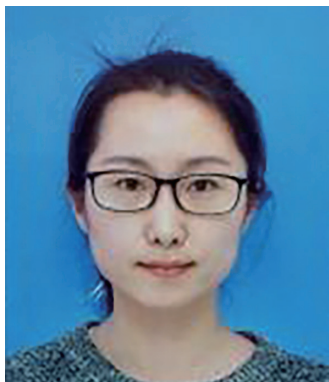
Conclusion

This chapter has introduced our method of metadata data similarity calculation in different languages and shown our method's ability to find the identical Ukiyo-e prints between Japanese and English databases. In the future, we will further improve the performance of this method by using other metadata fields. Currently, we only use the title and artist's name, but in the future, we want to use other metadata fields, such as the publication date. We also intend to apply our method to other languages, for example, between English and Dutch, and between Japanese and Dutch.

References

- Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2016. Learning Principled Bilingual Mappings of Word Embeddings While Preserving Monolingual Invariance. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2289–2294.
- Song, Yuting, Taisuke Kimura, Biligsaikhan Batjargal, and Akira Maeda. 2016. Proper Noun Recognition in Cross-Language Record Linkage by Exploiting Transliterated Words. *Proceedings of the 20th International Conference on Asian Language Processing (IALP 2016)*, 83–86.
- Song, Yuting, Biligsaikhan Batjargal, and Akira Maeda. 2017. Recognition and Transliteration of Proper Nouns in Cross-Language Record Linkage by Constructing Transliterated Word Pairs. *International Journal of Asian Language Processing*, 27(2), 111–125.
- . 2019a. Cross-Language Record Linkage based on Semantic Matching of Metadata. *DBSJ Journal*, 17(1), 1–8.
- . 2019b. Title Matching for Finding Identical Metadata Records in Different Languages. *Proceedings of the 13th International Conference on Metadata and Semantics Research (MTSR 2019)*, 431–437.

Dr. Yuting SONG



Chapter 3. Linking Ukiyo-e Records across Languages: An Application of Cross-Language Record Linkage Techniques to Digital Cultural Collections

Dr. Yuting Song is a Scientist at the Agency for Science, Technology and Research (A*STAR) in Singapore. Before joining A*STAR, she was a Specially Appointed Assistant Professor

at the College of Information Science and Engineering, Ritsumeikan University. She received her Ph.D. in Information Science and Engineering from Ritsumeikan University. Her research focuses range across natural language processing, cross-lingual information processing, digital libraries, and digital humanities. She is particularly interested in the potential of applying text and language processing technologies to the digital library and digital humanities domains to enhance accessibility, understanding, and utilization of vast collections of digital data.