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

Yuting SONG[*1], Biligsaikhan BATJARGAL[*2], Akira MAEDA[*3]

Abstract: Ukiyo-e is known worldwide as one of the traditional Japanese arts which flourished in the Edo period (1603–1868). There are many copies printed from the same ukiyo-e woodblocks, and they have been digitized and exhibited on the Internet by many cultural institutions around the world with descriptive metadata in various languages. This means that identical ukiyo-e prints can exist in different digital collections, along with metadata in different languages. This paper reports a summary of our work on cross-language record linkage, which is applied to find the identical ukiyo-e records across multiple data sources in different languages.

Keywords: *Japanese artworks, Cross-language record linkage, Digital cultural collections, Multilingual metadata.*

1. Introduction

With the development of digitization techniques, more and more cultural institutions, such as libraries, museums, and galleries, have been digitizing their cultural collections and exhibiting them online. It provides a new way to explore these cultural collections. Ukiyo-e is known worldwide as one of the traditional Japanese arts which flourished in the Edo period (1603–1868). Many cultural institutions in Japan and western countries hold ukiyo-e woodblock prints. Today, many of these prints have been digitized and exhibited on the Internet with metadata values (e.g., titles, artist names of ukiyo-e prints) in different languages.

[*1] Specially Appointed Assistant Professor, College of Information Science and Engineering, Ritsumeikan University
[*2] Senior Researcher, Kinugasa Research Organization, Ritsumeikan University
[*3] Professor, College of Information Science and Engineering, Ritsumeikan University
E-mail: [*1] songyt@fc.ritsumei.ac.jp
[*2] biligee@fc.ritsumei.ac.jp
[*3] amaeda@is.ritsumei.ac.jp
Published online: October 30, 2020.

Asia-Japan Research Academic Bulletin, 2019-2020.
ONLINE ISSN 2435-306X, Vol.1, 52.

As shown in Table 1, usually metadata values of original ukiyo-e prints are written in Japanese. However, in the organizations outside Japan, the metadata values are written in English or in the native language of that country.

Table 1. Some ukiyo-e print collections with metadata in different languages.

Collections	Location	Language	Total number
Ritsumeikan University Art Research Center	Japan	Japanese	approx. 12,000
Edo Tokyo Museum	Japan	Japanese	approx. 6,100
Waseda University Theatre Museum	Japan	Japanese	approx. 46,200
British Museum	United Kingdom	English	approx. 9,100
Metropolitan Museum of Art	United States	English	approx. 4,300
Rijksmuseum	Netherlands	Dutch	approx. 1,200
French Photo Agency	France	French	approx. 1,000

Figure 1. The examples of the identical ukiyo-e metadata records from different digital collections in Japanese and English.

This means that identical ukiyo-e prints can exist in different digital collections, along with metadata in different languages, as shown in Figure 1. Thus, language barriers are one of the challenges when people want to link or integrate ukiyo-e databases in different languages.

In this paper, we make a summary of our work on cross-language record linkage (Song *et al.*, 2017; Song *et al.*, 2019), which aim to identify the same records across multiple data sources in different languages. Moreover, we show the performance of applying our methods on linking the identical ukiyo-e prints.

To find the identical ukiyo-e records, Resig (2013) developed an ukiyo-e print search system[1], which can search identical ukiyo-e prints from multiple databases based on image similarity. Different from the image similarity-based method, our work focuses on finding identical ukiyo-e records by comparing their textual metadata values, including the titles and artist names of ukiyo-e prints. Our method can be easily applied to other databases that have no image resources, such as audio, video, 3D data and so on.

The remainder of this paper is structured as follows. Section 2 introduces the record linkage techniques and describes the general process of cross-language record linkage. Section 3 introduces our methods of metadata comparison across languages. Section 4 presents our evaluations on ukiyo-e print databases in Japanese and English. Section 5 concludes the paper and outlines future work.

[1] Ukiyo-e print search system: https://ukiyo-e.org

2. Cross-Language Record Linkage

(1) Record Linkage Techniques

Record linkage (Koudas *et al.*, 2006) is the task of identifying records that refer to the same entities from several data sources. It has been studied over the past decade and applied to the national census, health sector, and other application areas.

To identify the record pairs that refer to the same entity, the similarity between two records is calculated by comparing their metadata. The metadata of records contains different types of data, for example, the personal names, titles, and abstracts are string values, the financial data such as salaries and expenses are numerical values, and the date, age, and time, which are a special case of numerical values. This paper focuses on measuring the similarities of descriptive metadata (e.g., the titles of ukiyo-e prints), since the descriptive metadata given to an entity summarizes and distinguishes it from other entities.

(2) The Overall Process of Cross-Language Record Linkage

Figure 2 shows the general process of cross-language record linkage. Given the records in the source language and target language, record pairs are compared according to their metadata similarities. Then, based on these metadata similarities, the record pairs are classified into matches and non-matches by using a certain decision model. Finally, the matched record pairs are determined as the identical records that refer to the same real-world entities.

The source language is defined as the language of the records that are used to match the records in the other language. The other language is the target language. For example, we use the ukiyo-e metadata records in Japanese to match the ukiyo-e metadata records in English. In this case, Japanese is the source language, and English is the target language.

Our work mainly focuses on the part of metadata comparison, which measures the similarities between metadata in different languages.

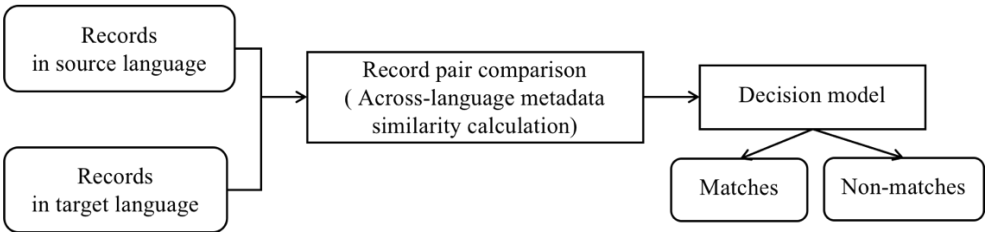

Figure 2. The general process of cross-language record linkage.

3. Across-Language Metadata Comparison

(1) Translation-based Method

To overcome language barriers, one solution is to translate the metadata in different languages into the same language by using machine translation. After the translation process, for each record in the source language, its translated metadata are compared with the metadata of records in the target

language. One problem that may arise is word mismatches between the translated metadata of the source language and metadata in the target language that describe the same objects. As shown in Figure 3, the word "夜" in the Japanese title is translated into "night" by using machine translation. However, it should be matched to "evening" in the corresponding English title.

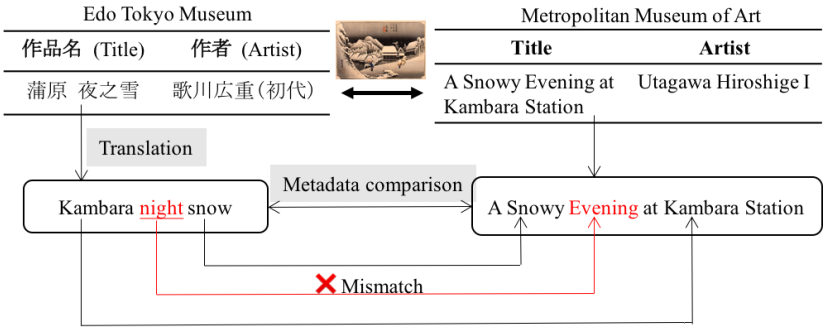

Figure 3. An example of word mismatches between translated metadata of the source language and its corresponding metadata in the target language.

Our proposed method (Song *et al.*, 2017) aims to deal with the mismatching problem between the translated metadata of source language and the metadata in the target language. In this method, word embeddings (Mikolov *et al.*, 2013) are employed to perform the semantic matching between the translated metadata of the source language and the metadata in the target language. Word embeddings are dense vector representations of words, which can better capture the semantic word relationships. By using such a property of word embeddings, the words in metadata are represented as vectors using word embeddings. In this way, two different words between translated metadata of the source language and metadata in the target language that express the same or similar meaning (e.g., the words "night" and "evening" in Figure 3) can be matched. Finally, the similarity between metadata is calculated as the cumulative maximum similarity that the words in the translated metadata match the words in metadata in the target language.

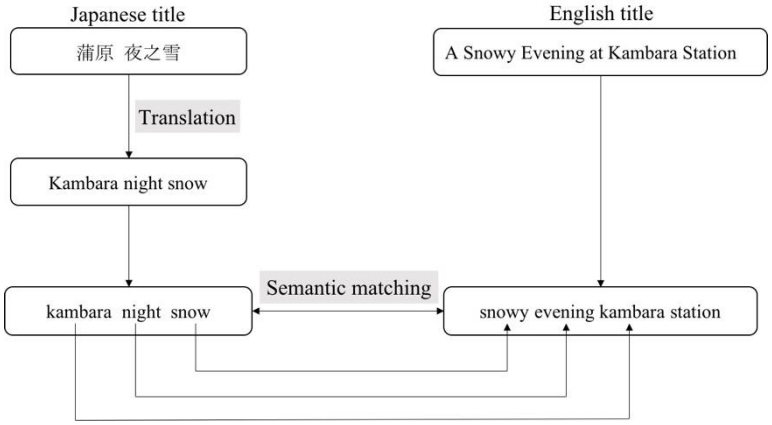

Figure 4. An example that illustrates our translation-based method.

Figure 4 illustrates our method to calculate the similarity between ukiyo-e titles in Japanese and English. First, the Japanese title is translated into English by using machine translation. Then, English stopwords (e.g. "a" and "at" in the English title) are removed, and all the words are converted to lowercase. Next, each word in the translated title is used to match all the words in the English title by using word embeddings. In Figure 4, the arrow from the word in the translated title to the word in the

English title represents they are matched with the maximum similarity score. Since the word “night” in the translated title and “evening” in the English title has a semantically similar meaning, they can be matched by using word embeddings.

(2) Bilingual Word Embedding Based Method

We also attempt another way to calculate the similarities of metadata in different languages by using bilingual word embeddings (BiWE)[2], which is a method without relying on translation. Bilingual word embeddings (Conneau *et al.*, 2018) are cross-lingual vector representations of words, which leverage the same vector space for two different languages so that similar words in different languages can have similar vector representations. As shown in Figure 5, cross-language similar words such as “世界” (“world” in Japanese) and “world” have close vectors. While word embedding vectors are usually hundreds of dimensions (e.g., two or three hundred), only two dimensions are shown in Figure 5 for simplicity.

We proposed a method of measuring the metadata similarity by using bilingual word embeddings (Song *et al.*, 2019). This method calculates metadata similarities through word-to-word matching between the titles in different languages. Let b be the vector space of bilingual word embeddings. Each word w in the source language and target language can be represented by the bilingual word embeddings as $b(w)$. Each word in the metadata in the source language is used to calculate the cosine similarity with all the words in the metadata in the target language, and the maximum similarity score is used as the contribution of this word to the metadata similarity, which is formulated in equation (1).

$$\text{Metadata similarity} = \sum_{i=1}^{N_{t_S}} \max_{w_j^{t_T} \in t_T} cosine\left(b(w_i^{t_S}), b\left(w_j^{t_T}\right)\right) \quad (1)$$

where t_S is the metadata in the source language, $w_i^{t_S}$ is word i in t_S, N_{t_S} is the number of words in t_S; t_T is the metadata in the target language, $w_j^{t_T}$ is the word j in t_T.

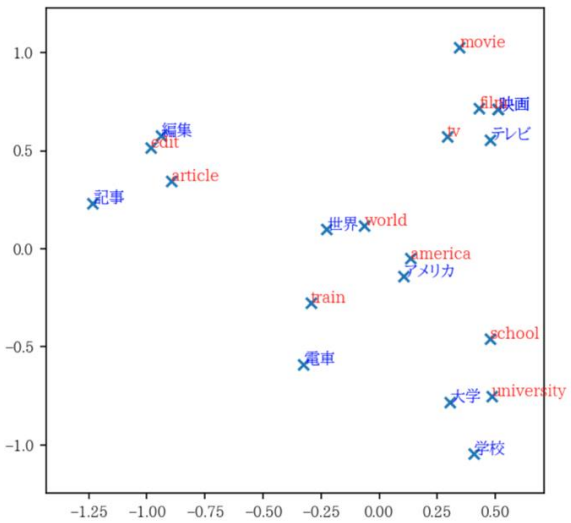

Figure 5. Japanese-English word embedding space.

[2] Bilingual word embeddings can be trained using the toolkit: https://github.com/facebookresearch/MUSE. The pretrained bilingual word embeddings can be downloaded from https://fasttext.cc/docs/en/aligned-vectors.html

4. Evaluation on Ukiyo-e Prints Datasets in Different Languages

(1) Experimental Setup

We applied our work to the datasets of ukiyo-e metadata records in Japanese and English, which is shown in Table 2. Each Japanese ukiyo-e metadata record in the Japanese dataset has at least one corresponding ukiyo-e metadata record in the English dataset, which means they refer to the identical ukiyo-e prints.

Table 2. Experimental datasets of ukiyo-e metadata records in Japanese and English.

Language	Databases	Number of ukiyo-e records
Japanese	Edo Tokyo Museum	203
English	Metropolitan Museum of Art	3,398

We used the titles, series names 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 the artists of ukiyo-e prints.

For the translation based method, we used the Microsoft Translator Text API[3] to translate Japanese metadata into English. As it provides two translation models: statistical machine translation (SMT) and neural machine translation (NMT), we experimented with both translation models. After translation, we used two methods for metadata similarity calculation: 1) soft-tfidf similarity (MT-soft-tfidf): it is a string-based similarity metric and showed the best performance in title matching against 20 other commonly used string-based similarity metrics (Gali *et al.*, 2017); 2) semantic matching (MT-semantic): This is our proposed method based on monolingual word embeddings (Song *et al.*, 2017).

(2) Results and Analysis

Figure 6 shows the experimental results. Experimental results are evaluated by top-k precision, which is defined as the percentage of Japanese titles that have correct corresponding English titles in its top-k candidates. The higher the values are, the better the given method works.

a) Machine translation vs. Bilingual word embeddings. The method that relies on machine translation generally outperforms the method that relies on bilingual word embeddings. It can be seen that the performance of BiWE method is higher on Top-1 precision than MT-soft-tfidf (SMT). This is an encouraging finding because bilingual word embeddings can be easily applied to other low resource language pairs, in which it is difficult to find a large amount of bilingual parallel data for training the MT system.

b) SMT vs. NMT. NMT, as the current state-of-the-art model for the machine translation system, has been proved to have a more precise translation quality than SMT. We can also see that the results of the machine translation method of using NMT are better than SMT, which indicates that the performance of the machine translation based method is influenced by the translation quality.

[3] Microsoft Translator Text API: https://www.microsoft.com/en-us/translator/translatorapi.aspx

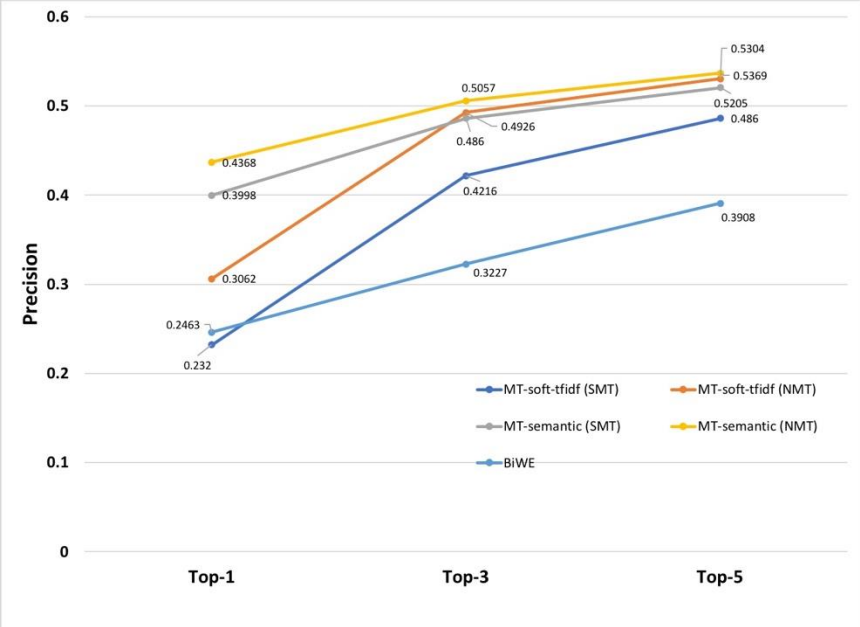

Figure 6. Experimental results: Top-k precision.

5. Conclusions

In this paper, we report a summary of our work on cross-language record linkage. Moreover, we showed the performance of applying our methods on linking ukiyo-e print databases in Japanese and English. In the future, we plan to apply our work to the ukiyo-e prints in other languages, such as Dutch. We also plan to apply our methods to other digital cultural collections.

References

Conneau, A., Lample, G., Ranzato, M.A., Denoyer, L. and Jégou, H. 2018. *Word translation without parallel data*. Proc. of the Sixth International Conference on Learning Representations.

Koudas, N., Sarawagi, S., Srivastava, D. 2006. *Record linkage: Similarity Measures and Algorithms.* Proc. of the 2006 ACM SIGMOD International Conference on Management of Data. pp. 802–803.

Gali, N., Mariescu-Istodor, R. and Fränti, P. 2017. *Similarity Measures for Title Matching.* Proc. of 23rd International Conference on Pattern Recognition, pp. 1549-1554.

Song, Y., Kimura, T., Batjargal, B. and Maeda, A. 2017. *Cross-Language Record Linkage by Exploiting Semantic Matching of Textual Metadata*. Proc. of the 9th Forum of Data Engineering and Information Management in Japan (DEIM 2017).

Song, Y., Batjargal, B. and Maeda, A. 2019. *Title Matching for Finding Identical Metadata Records in Different Languages*. Proc. of the 13th International Conference on Metadata and Semantics Research (MTSR2019), pp. 431-437.

Resig, J. 2013. *Aggregating and analyzing digitized Japanese woodblock prints.* In 3rd Annual Conference of the Japanese Association for Digital Humanities. Kyoto, Japan. https://ukiyo-e.org/about.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. 2013. *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781.