Cross-Language Record Linkage Across Humanities Collections Using Metadata Similarities Among Languages

Yuting Song(✉)

Ritsumeikan University, Kusatsu, Shiga 5258577, Japan
gr0260ff@ed.ritsumei.ac.jp

Abstract. This paper proposes a method for cross-language record linkage across digital humanities collections by exploiting similarities between metadata values in different languages without using any translation method. Our method represents metadata values in Japanese and English as vectors by using monolingual word embeddings. Then, we calculate similarity between metadata value vectors by learning a mapping between vector spaces that represent Japanese and English. The proposed method could help users to acquire multilingual information of the objects in digital collections. We evaluate the effectiveness of our method on Japanese Ukiyo-e print databases in Japanese and English.

Keywords: Cross-language record linkage · Word embeddings · Digital humanities collections

1 Introduction

Over the past decade, more and more libraries, museums and galleries around the world have been digitalizing their collections and making them accessible online. It opens up new opportunities to acquire valuable knowledge from vast amounts of information about these digital collections. The metadata, which are used to provide information about the records in digital collections, are created independently by heterogeneous institutions using different natural languages. For instance, Japanese Ukiyo-e woodblock prints[1] have been digitized by many museums in Japan and Western countries and described by the metadata values in their native languages. As a consequence, identical records that refer to the same object could be described in different languages. Given that there is multilingual information in metadata of identical records, it is important to provide technologies for finding these identical records in order to aggregate multilingual knowledge about objects.

Record linkage [1] is a task of finding record pairs that refer to the same object across multiple data sources, which has been studied for many years. Our research focuses on a new field of cross-language record linkage, where records are from the data sources with metadata in different languages. In particular, we aim for cross-language record linkage across digital humanities collections in Japanese and English by using textual metadata

[1] Ukiyo-e is a type of Japanese traditional woodblock print, which is known as one of the popular arts of the Edo period (1603–1868).

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 640–643, 2017.
DOI: 10.1007/978-3-319-67008-9_60

values. It is challenging due to several reasons: (1) metadata values are expressed in different languages, therefore similarity measures cannot be employed directly; (2) even if the machine translation system can be used to translate metadata values into the same language in order to calculate similarities between them, machine translation systems have poor performance on specific domains [2] due to the difficulty of obtaining a domain-specific bilingual corpus for training system.

2 Proposed Method

Our proposed method focuses on the similarity matching phase of cross-language record linkage, which is an important phase that determines whether two records represent the same object. In this section, we first introduce our approach of representing textual metadata values. Then, a method of learning a mapping between vector spaces that represent Japanese and English is provided for calculating the similarity between metadata values.

2.1 Representations of Metadata Values

We represent textual metadata values as vectors by using word embeddings [3], which are dense, low-dimensional and real-valued vectors for representing words. Through these embedded word representations, the words with a similar meaning have closer distances in a vector space, e.g. *vector*("storm") is close to *vector*("hurricane"), which means the semantic relationships between words can be captured. Moreover, the semantic relationships between words can be expressed as linear operations in a vector space, e.g., *vector*("Berlin") - *vector*("Germany") + *vector*("France") is close to *vector*("Paris").

Our method of representing textual metadata values is inspired by the characteristics of word embeddings. More specifically, we firstly learn Japanese and English word embeddings by using Word2Vec toolkit. Then, we represent textual metadata values in Japanese and English by additive combination of the vector embeddings of words that compose the metadata values. Fig. 1 illustrates our method of representing metadata values.

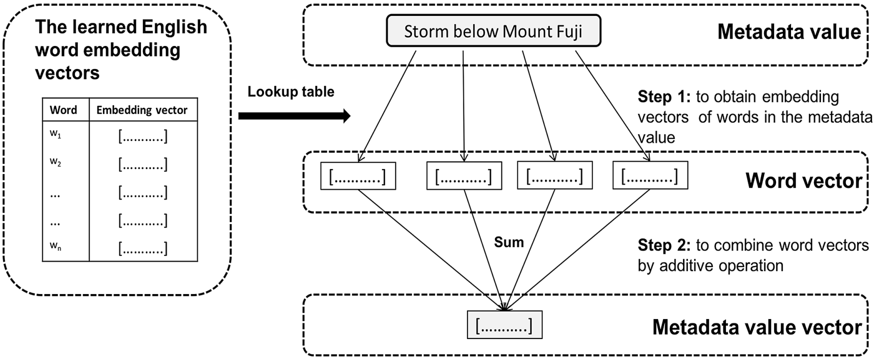
The learned English word embedding vectors
Word
Embedding vector
w1
w2
...
...
wn
[..........]
Lookup table
Storm below Mount Fuji
Metadata value
Step 1: to obtain embedding vectors of words in the metadata value
Word vector
Sum
Step 2: to combine word vectors by additive operation
Metadata value vector

Fig. 1. The process of our method of representing metadata values

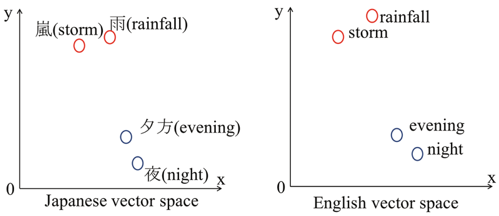

Fig. 2. The word vector representations of weathers and times in Japanese and English

2.2 Similarity Calculation Between Metadata Value Vectors

We calculate the similarity between metadata value vectors by learning a mapping between vector spaces that represent Japanese and English. Our proposed method is motivated by the idea in [4] that the same concepts have similar geometric arrangements across the vector spaces that represent different languages, which is illustrated in Fig. 2. Taking the concept of weather as an example, the relative positions of “雨 (rainfall)” and “嵐 (storm)” in the vector space that represents Japanese (the graph on the left) are similar to the relative positions of “rainfall” and “storm” in the English vector space (the graph on the right). What is more important is that the relationship between vector spaces that represent these two languages can possibly be captured by learning a mapping between them, e.g. a liner mapping. If we know some word pairs in Japanese and English, e.g. “雨” and “rainfall”, “嵐” and “storm”, we can learn a mapping that can help us to transform other words in the Japanese vector space to the English vector space.

Similar to the idea above, we learn a liner mapping between vector spaces that represent Japanese and English in order to transform the Japanese metadata value vectors to the vector space that represents English. Suppose we are given a set of textual metadata value pairs and their associated vector representations $\{x_i, z_i\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^1$ is the vector representation of Japanese metadata value i, and $z_i \in \mathbb{R}^2$ is the vector representation of its corresponding English metadata value that is obtained by our method in Sect. 2.1. Our goal is to find a mapping matrix W such that Wx_i approximates z_i. In practice, W can be learned by the following optimization problem shown in Eq. (1), which can be solved with stochastic gradient descent.

$$\min_W \sum\nolimits_{i=1}^{n} \| Wx_i - z_i \|^2 \tag{1}$$

At the time of similarity calculation, for any given new Japanese metadata value vector x, we transform it into a vector space that represents English by computing $z = Wx$. Then, we can calculate the similarity between metadata values in Japanese and English by comparing the transformed vectors of Japanese metadata with other English metadata value vectors.

3 Experiment

In this section, we show the preliminary results of our proposed method in finding the identical Ukiyo-e prints across databases in Japanese and English.

In the experiments, the titles of Ukiyo-e prints are used to calculate similarities between Ukiyo-e prints. We train the Japanese and English word vectors on Japanese and English Wikipedia articles using Word2Vec toolkit. In the process of learning the mapping between the vector spaces that represent Japanese and English, we use 600 Japanese-English parallel short sentence pairs for pre-training. In order to make this mapping more accurate to transform Ukiyo-e titles, we further use 74 pairs of Japanese and English Ukiyo-e titles to optimize this mapping, in which each pair of titles refers to the same Ukiyo-e prints. The similarities between the Ukiyo-e titles are calculated by cosine similarity metric.

We use 173 pairs of Japanese and English Ukiyo-e titles as the test data to evaluate our method. The precision at top-n are used to evaluate the experimental results. In order to verify the effectiveness of using Ukiyo-e titles to optimize the mapping in the phase of pre-training, we show the results of both conditions of using Ukiyo-e titles and without using them. The experimental results are shown in Table 1.

These results show that the precisions can be improved by using Japanese and English Ukiyo-e titles to optimize the mapping between Japanese and English vector spaces. The experimental results also confirm the usefulness of our proposed method for linking the same Ukiyo-e prints in Japanese and English.

Table 1. The experimental results

	Precision			
	In top-1	Within top-5	Within top-10	Within top-15
Without using Ukiyo-e titles	2.3%	12.2%	17.4%	22.7%
Using Ukiyo-e titles	29.1%	41.9%	50.0%	54.7%

4 Conclusions

In this paper, we proposed a method of cross-language record linkage by measuring the similarity between textual metadata values without using any translation methods. In the future, we plan to explore different embedded vector representations of metadata values. Besides we will evaluate our method on data sources in other languages.

References

1. Ahmeh, K.E., Panagiotis, G.I., Vassillios, S.V.: Duplicate record detection:a survey. IEEE Trans. Knowl. Data Eng. **9**(1), 1–16 (2007). IEEE
2. Hua, W., Haifeng, W., Chengqing, Z.: Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In: 22nd International Conference on Computational Linguistics, pp. 993–1000. ACL, Manchester (2008)
3. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
4. Mikolov, T., Le, Q.V. and Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013)