Cross-Language Record Linkage using Word Embedding driven Metadata Similarity Measurement

Yuting Song[1], Taisuke Kimura[1], Biligsaikhan Batjargal[2], Akira Maeda[3]

[1]Graduate School of Information Science and Engineering, Ritsumeikan University, Japan
{gr0260ff, is0013hh}@ed.ritsumei.ac.jp
[2]Research Organization of Science and Engineering, Ritsumeikan University, Japan
biligee@fc.ritsumei.ac.jp
[3]College of Information Science and Engineering, Ritsumeikan University, Japan
amaeda@is.ritsumei.ac.jp

Abstract. Aiming to link the records that refer to the same entity across multiple databases in different languages, we address the mismatches of wordings between literal translations of metadata in source language and metadata in target language, which cannot be calculated by string-based measures. In this paper, we propose a method based on word embedding, which can capture the semantic similarity relationships among words. The effectiveness of this method is confirmed in linking the same records between Ukiyo-e (Japanese woodblock printing) databases in Japanese and English. This method could be applied to other languages since it makes little assumption about languages.

Keywords: Cross-language record linkage ·Similarity measurement ·Word embedding ·Semantic matching

1 Introduction

Cross-language record linkage is a task of finding pairs of records that refer to the same entity across multiple databases in different languages. It is crucial to various fields, such as federated search and data integration. Furthermore, the metadata of identical records in different languages are helpful to build multilingual Linked Data. Cross-language record linkage consists of two steps. First, the metadata of a record, e.g. title, author, publisher, in the source language are translated into the target language based on bilingual dictionaries. Next, identical records are determined by calculating the similarities between metadata within one language, which is similar to the monolingual record linkage [1].

In monolingual record linkage, the mismatches are mainly due to the typographical variations of string data, which can be measured by string-based comparison. Nevertheless, when it comes to cross-language record linkage, the mismatches of wordings between literal translations and metadata in target language cannot be measured by simple metrics. Figure 1 gives an example of this type of mismatch. The word "白雨" in Japanese is translated into "rainfall" by a Japanese-English bilingual dictionary.

However, the corresponding word in English title is “storm”, which is translated by a human expert translator. Such a mismatch is due to the use of different wordings to express the same meaning, which cannot be measured by string-based similarity. Some approaches exploit the network structure of records deeply in knowledge bases to determine the identical records [2]. However, in most databases, unlike Wikipedia or WordNet, the network structure of records cannot be obtained easily.

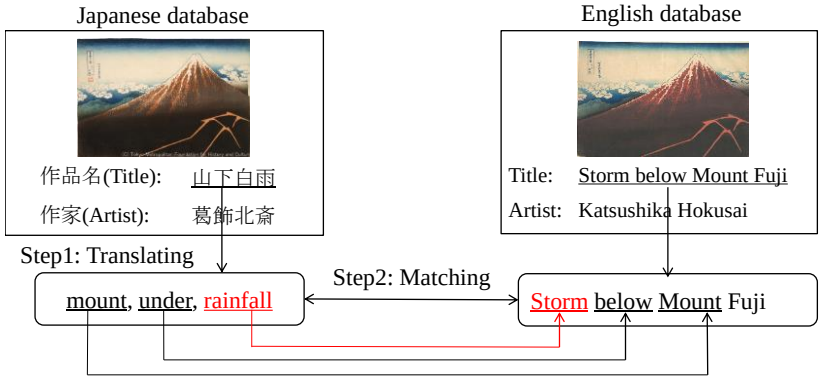

Fig. 1. An example of mismatches of wordings between literal translations of metadata in source language and metadata in target language

In this paper, we propose a method for cross-language record linkage that can measure the similarities between metadata with the same meaning but in different wordings. Our method is based on distributed representations of words [3] (a.k.a. word embedding), in which semantically similar words are closer in vector space. The effectiveness of this approach is evaluated in the record linkage across Ukiyo-e databases in Japanese and English.

2 Methodology

As mentioned above, cross-language record linkage can be divided into two steps: translating and matching. We focus on the second step, especially the matching among non-proper nouns in metadata. The reason is that non-proper nouns are more likely to be translated into different words than proper nouns. Proper nouns can usually be transliterated, which have a one-to-one mapping.

2.1 Learning Distributed Representations of Words

Distributed representations for words are dense, low-dimensional and real-valued vectors, which were firstly proposed by Rumelhart et al. [4]. Recently, the distributed skip-gram model for learning word representations was introduced by Mikolov et al. [3]. This model employs simple neural network architecture, which can be trained on a large amount of unstructured text data in a short time (billions of words in hours). Besides, the distributed representations of words learnt by this model can capture

semantic similarity relationships. Considering the advantages above, we utilize the skip-gram model of Mikolov et al. for learning word representations in our method.

2.2 Similarity Measurement between Metadata

In the proposed method, the similarity metric between the literal translations of metadata in source language (M_{lt}) and metadata in target language (M_t) is defined in Formula 1. $NP(M_{lt})$, $NP(M_t)$ are the number of non-proper nouns in M_{lt} and M_t respectively. np_i is a non-proper noun in M_{lt}. $C(np_i)$ is the number of candidate translations of np_i . v_{ij} is the distributed representation of a candidate translation of np_i. Similarly, v_q is the distributed representation of a non-proper noun in M_t. $score(np_i)$ is the matching degree of np_i, which is the maximal value of similarity between candidate translations of np_i and non-proper nouns in M_t. $cosine(v_{ij}, v_q)$ is the cosine similarity between v_{ij} and v_q. N_p means the number of matched proper nouns. w_p and w_{np} are weights of proper nouns and non-proper nouns respectively. L is the total number of words in M_{lt}.

$$S(M_{lt}, M_t) = [\, w_p \cdot N_p + w_{np} \cdot \sum_{i=1}^{NP(M_{lt})} score(np_i) \,]/L \quad (1)$$

where $score(np_i) = \max\ [\sum_{j=1}^{C(np_i)} \sum_{q=1}^{NP(M_t)} cosine(v_{ij}, v_q)\,]$

3 Experiments

In this section, we evaluate the effectiveness of our proposed method in linking the same Ukiyo-e prints between the databases in Japanese and English.

3.1 Experimental Setup

The titles of Ukiyo-e prints are used to identify the same records. The experimental data set consists of 243 Japanese titles of Ukiyo-e prints in the Edo-Tokyo Museum[1] and 3,293 English titles in the Metropolitan Museum of Art[2], in which each Japanese title has at least one corresponding English title. Among the 243 Japanese titles, 143 titles are descriptive titles that contain at least one non-proper noun.

Here we translate non-proper nouns of Japanese titles into English by using EDR Japanese-English bilingual dictionary[3]. The proper nouns are transliterated by Hepburn Romanization system[4]. Distributed representations of words are learnt from the text data in English Wikipedia dump that contains more than 3 billion words. The similarities between the literal translations of Japanese titles and English titles are calculated by our proposed method (Formula 1). Besides, we use a baseline for com-

[1] http://digitalmuseum.rekibun.or.jp/app/selected/edo-tokyo
[2] http://www.metmuseum.org/
[3] http://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html
[4] https://en.wikipedia.org/wiki/Hepburn_romanization

parison experiments. It is using string matching to measure the similarities among words in titles [5], which is shown in Formula 2. N_p and N_{np} are the number of matched proper nouns and non-proper nouns in literal translations of Japanese titles respectively. w_p and w_{np} are their weights. L is the total number of words in a Japanese title. We set w_p, w_{np} equal to 2 and 1 respectively, which is the same as [5]. Here, proper nouns are given a higher weight than non-proper nouns, because proper nouns are representative features for calculating similarity in our proposed method.

$$\text{Similarity metric} = (w_p \cdot N_p + w_{np} \cdot N_{np})/L \quad (2)$$

3.2 Experimental Results

Table 1 shows the performance of baseline and our proposed method for cross-language record linkage using descriptive titles and all titles. From the results, it can be seen that our proposed method is better than the baseline method, especially for descriptive titles that contain one or more non-proper nouns.

Table 1. Results of cross-language record linkage.

	The precision of descriptive titles	**The precision of all titles**
Baseline	0.31	0.27
Our method	0.43	0.34

4 Conclusion

In this paper, we proposed a method that employs the distributed representations of words to measure metadata similarities for cross-language record linkage. Experimental results have shown that this approach improves the precision of cross-language record linkage between Ukiyo-e databases in Japanese and English. In the future, we plan to improve the similarity metric by measuring the degree of similarity between word embedding.

References

1. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering, 19(1), 1-16 (2007)
2. Pilehvar, M.T., Navigli.R.:A Robust Approach to Aligning Heterogeneous Lexical Resources. In *ACL*, 468-478(2014)
3. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781. (2013)
4. Rumelhnrt, D.E., Hinton, G.E., Williams, R.J.: Learning Representations by Back-Propagating Errors. Nature. 323,533-536 (1986)
5. Kimura, T., Batjargal, B., Kimura, F., Maeda, A.: Finding the Same Artworks from Multiple Databases in Different Languages. In Conference Abstracts of Digital Humanities (2015)