Regular Paper

DBSJ Journal, Vol. 17 No. 1
March 2019

Cross-Language Record Linkage based on Semantic Matching of Metadata

Yuting SONG♡
Biligsaikhan BATJARGAL◇
Akira MAEDA♠

Record linkage is finding record pairs that refer to the same entities or objects across multiple data sources. This task is crucial in various research fields, such as federated search and data integration. This article focuses on the new challenge of cross-language record linkage, where records are from data sources in different languages. To compare the records in different languages, the records' metadata needs to be translated from a source language to a target language. This causes mismatches between the translated metadata and the metadata in the target language, since similar meanings are sometimes expressed by different words during translation. Thus, conventional string-based similarity metrics are insufficient for measuring the similarities between the translated metadata and the metadata in the target language. Therefore, we propose a method of dealing with the mismatching problem in cross-language record linkage. For each translated metadata of the source language, first, we use a string-based similarity metric to identify the potential matching metadata in the target language as candidates. Then, we employ word embeddings to perform the semantic matching between the translated metadata and its candidate matching metadata in the target language. Our method is evaluated on a real-world dataset in Japanese and English. Our experiments proved that our proposed method outperforms baseline methods that only rely on string similarities or the semantic matching method.

1. Introduction

Record linkage [1][2] is finding record pairs that refer to the same entities or objects across different data sources. It is also known as object identification [3] or duplicate detection [4]. To determine whether a pair of records refer to the same entity or not, the records from one data source need to be compared with records in other data sources according to their metadata similarities. With the emergence of multilingual data sources, record linkage techniques should be able to work with records that are represented with metadata in different languages.

In this work, we focus on cross-language record linkage, which identifies record pairs that refer to the same entities or objects across different data sources in different languages. It can be used to expand and enrich the metadata of records in other languages. It can also help people acquire the metadata of a specific record regardless of the language.

An example application of cross-language record linkage involves linking the record pairs that refer to the same ukiyo-e prints in Japanese and English databases. Ukiyo-e is a type of Japanese traditional woodblock printing. There are many copies printed from the same ukiyo-e woodblocks, and they have been digitized and exhibited on the internet by many libraries and museums all around world with metadata in various languages [5]. For example, Figure 1 shows that the metadata of ukiyo-e prints in the Edo-Tokyo Museum is in Japanese. The metadata of the ukiyo-e prints at the Metropolitan Museum of Art is in English. Linking these sources requires comparing the metadata of records from both sources and identifying the record pairs that refer to the same ukiyo-e prints. Due to language barriers, the metadata cannot be compared directly. Therefore, the metadata in the source language needs to be translated into the target language in order to identify whether they refer to the same object or not.

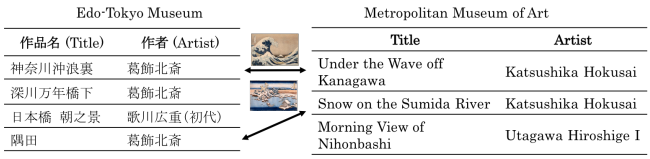

Figure 1: Matching of ukiyo-e prints

One problem that can arise when comparing metadata after translation is word mismatches between the translated metadata of the source language and metadata in the target language that describe the same objects. An example of such mismatches is shown in Figure 2. The word "夜" in the Japanese title is translated into "night" by Bing Microsoft Translator. However, the word that should be matched with "night" in the corresponding English title is "evening".

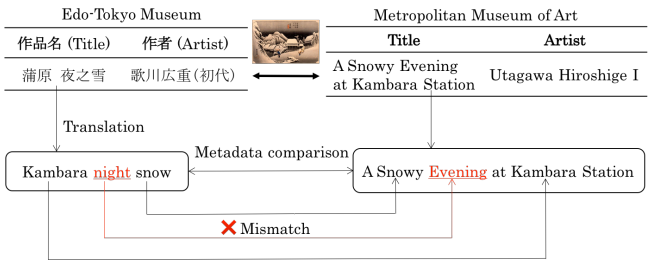

Figure 2: An example of word mismatches between translated metadata and metadata in the target language

♡Student member, Graduate School of Information Science and Engineering, Ritsumeikan University
aprilsongrits@gmail.com
◇Non-member, Kinugasa Research Organization, Ritsumeikan University
biligee@fc.ritsumei.ac.jp
♠Member, College of Information Science and Engineering, Ritsumeikan University
amaeda@is.ritsumei.ac.jp

In monolingual record linkage, mismatches between the metadata that describe the same object occur mainly due to the typographical variations of metadata. For example:

- Massey – Massie
- Peter Christen – Christian Pedro
- Ruby Lotel – Ruby L'otel

Many similarity measures are used to compare records' metadata in monolingual record linkage, such as edit distance based string comparison [6][7][8], q-gram based string comparison [9][10], Jaro and Winkler distance [11][12], and the Monge-Elkan method [13]. Such traditional string-based similarity measures—which are used in monolingual record linkage—impact metadata similarity calculation positively, especially for metadata that contain a word that refers to a particular thing, such as a proper noun. The translations of these words are usually in the same textual representation, unlike other words whose translations are sometimes synonyms. For example, the translation for the Japanese word "蒲原" in Figure 2 is "kambara", but "夜" is translated into "night" or "evening".

However, these traditional string-based similarity metrics are insufficient in determining whether two metadata are semantically similar in cross-language record linkage. This is because, when metadata in the source language are translated into the target language, the mismatches between metadata may occur due to different words that express similar meanings. For example, in Figure 2, the words "night" and "evening" are used to express a similar meaning.

Therefore, we propose a method of dealing with mismatching between metadata in cross-language record linkage. Specifically, our method focuses on descriptive metadata such as titles and abstracts, rather than metadata such as authors, dates, and formats. Since descriptive metadata summarizes the content of an entity or distinguishes it from other entities, it is more likely to be translated into different words that might lead to mismatches between metadata.

In our method, for each translated metadata of the source language, we first identify the candidate metadata in the target language using a string-based similarity metric. Then, we employ word embeddings to perform semantic matching between the translated metadata of the source language and the metadata in the target language. Our method leverages successful achievements in word embeddings [14], which is dense vector representations of words. The learned word embeddings demonstrate that they can better capture the semantic word relationships, which means semantically similar words are close in the vector space. By using this property of word embeddings, we represent the words in metadata as vectors using word embeddings. In this way, two different words between the translated metadata and the metadata in the target language that express similar meanings (e.g., the words "night" and "evening" in Figure 2) can be matched, since their embedding vectors are close in the vector space. Finally, the similarity between metadata is the maximum cumulative similarity, that the words in the translated metadata match the words in metadata in the target language. We evaluate our proposed method on the dataset in Japanese and English. Our experiments showed that our method improves the performance of cross-language record linkage compared to the baseline method that is based on string comparison. We also compared our method with our previous work [15], which directly used word embeddings to perform semantic matching between the translated metadata of the source language and the metadata in the target language. Our experiments showed that our method works better.

The remainder of the article is structured as follows. Section 2 describes the general process of cross-language record linkage. Section 3 introduces our method. Section 4 presents our experimental setup and evaluations. Section 5 outlines related work. The Conclusion concludes the paper and outlines future work.

2. Cross-Language Record Linkage

Figure 3 shows the general process of cross-language record linkage. First, records' metadata in the source language are translated into the target language in order to compare the metadata within the same language. The dictionary-based method and machine translation based method are commonly used in cross-language tasks (e.g., cross-language information retrieval and cross-language plagiarism detection)[16][17][18]. Record pairs are compared according to their metadata similarities after translation.

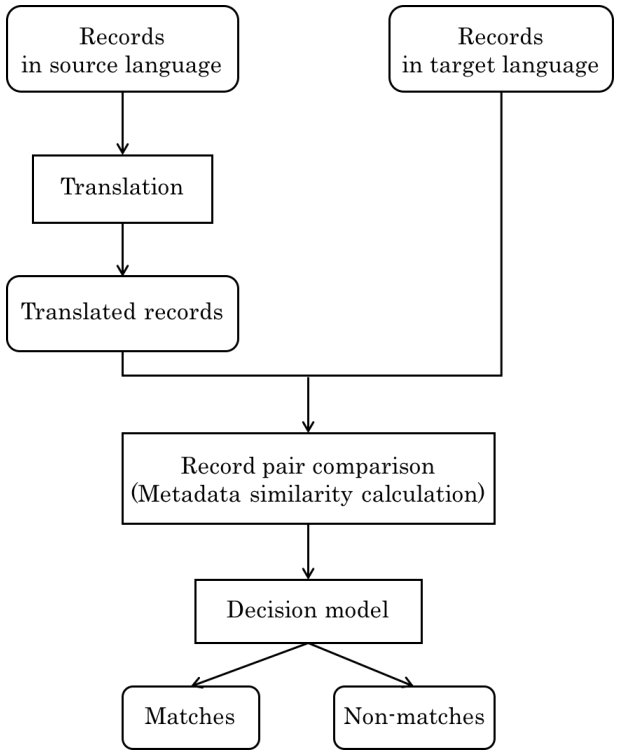

Figure 3: The general process of cross-language record linkage

Our work focuses on the second part, measuring the similarities between translated metadata and metadata in a target language. Finally, based on these metadata similarities, the record pairs are classified into matches and non-matches by using a certain decision model. The matched record pairs are determined as the identical records that refer to the same real-world entities.

3. Semantic matching of metadata

In this section, first, we introduce the idea of word embeddings, which are employed in our method to capture the semantic similarity between words. Then, we present our method, which consists of two steps: 1) candidate identification and 2) semantic matching.

3.1 Word embeddings

Word embeddings, which are distributed representations for words, were first proposed by Rumelhart et al. [19] and have achieved impressive results in many natural language processing tasks [20], such as parsing [21] and named entity recognition [22]. Mikolov et al. [14] introduced *word2vec*, which is a toolkit for learning word embeddings. It includes two word embedding models, the skip-gram and the continuous bag-of-words. These models learn word representations by employing simple neural network architecture. Specifically, the skip-gram model consists of three layers (input, projection, and output) to predict contextual words of the input word vector. The objective of training is to learn word vector representations that are good at predicting its context in the same sentence. Due to its simple architecture, the skip-gram model can be trained on a large amount of unstructured text data in a short amount of time (billions of words in a few hours) using a conventional desktop computer.

The main advantage of learned word vector representations is that semantically similar words are close in the vector space. We utilized *word2vec* to learn word embeddings. Other word embedding models, such as Glove [23], were also taken into consideration.

3.2 Candidate identification

To avoid unnecessary comparisons between metadata, we aim to obtain all the possible matched metadata in the target language for each translated metadata of the source language. The output of this step is taken as the input of semantic matching.

As we mentioned in Section 1, in cross-language record linkage, string-based similarity measures are insufficient for measuring the similarity between the translated metadata and the metadata in the target language, but we should not ignore their effectiveness in metadata similarity calculation. The key idea of candidate identification is that the translated metadata and the metadata in the target language are more likely to be similar if they share the same words.

Thus, we identify the matched candidate metadata based on whether a metadata in the target language shares the same words with the translated metadata. Given a translated metadata (M_{trans}) and a metadata in the target language (M_{target}), each of them is segmented into a set of words. After removing stop words, M_{trans} and M_{target} contain a set of words W_{trans} and W_{target}, respectively. The string-based similarity (SS) between M_{trans} and M_{target} is defined as:

$$SS(M_{trans}, M_{target}) = \begin{cases} 1, & \text{if } W_{\text{trans}} \cap W_{target} \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

If $SS(M_{trans}, M_{target})$ equals 1, M_{target} will be identified as one of the matched candidate metadata of M_{trans}. Figure 4 shows why the English title "A Snowy Evening at Kambara Station" is identified as the candidate title.

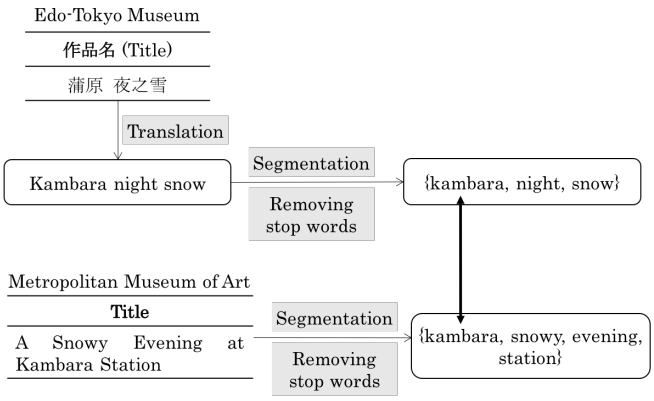

Figure 4: An example of candidate identification

3.3 Semantic matching

After identifying candidates, we perform semantic matching between each translated metadata and its matched candidate metadata in the target language by employing word embeddings.

Assume we are provided with a word embedding matrix $W \in R^{n \times d}$ for a finite size vocabulary of n words. The i-th row, $w_i \in R^d$, represents the embedding of the i-th word. The dimension of the word embedding space is d.

Our proposed method incorporates the semantic similarity between word pairs into the similarity between the translated metadata and metadata in target language. Here, we use a cosine similarity metric to measure the word similarities. Specifically, the semantic similarity between word i and word j is shown in Equation (2).

$$word_{sim}(i, j) = cosine(w_i, w_j) \quad (2)$$

Our goal is to measure the semantic similarity between the translated metadata and metadata in the target language. Intuitively, if translated metadata contains more words that can match the words in metadata in the target language, either exactly or semantically, they might more possibly describe an identical entity. We represent metadata as a set of embedded words. The similarity between the translated metadata (M_{trans}) and

metadata in the target language (M_{target}) is formulated as the cumulative similarity of word pairs between M_{trans} and M_{target}. First, for each word i (w_i) in M_{trans}, we calculate the similarities between w_i and each word j (w_j) in M_{target}. The similarity between w_i and w_j is calculated by Equation (1). Then, the maximum similarity between w_i and each word j (w_j) in M_{target} is regarded as the similarity contribution of w_i to the similarity between M_{trans} and M_{target}, which is shown in Equation (3).

$$Contri(w_i) = \max\big(word_{sim}(i,j)\big) \ \forall j \in \{1, \dots, n\} \quad (3)$$

$Contri(w_i)$ represents the similarity contribution of w_i in M_{trans} to the similarity between M_{trans} and M_{target}. n represents the number of words in M_{target}.

Finally, we can define the similarity between M_{trans} and M_{target} as the cumulative similarity contribution of each w_i in M_{trans}, which is shown in Equation (4).

$$Sim\big(M_{trans}, M_{target}\big) = \frac{1}{m}\sum_i^m Contri(w_i) \quad (4)$$

$Sim\big(M_{trans}, M_{target}\big)$ represents the similarity between M_{trans} and M_{target}. m represents the number of words in M_{trans}.

Our proposed method is named as SS+SM, since it identifies the candidates by using the string-based similarity (SS) metric, which is defined in Equation (1), and it utilizes semantic matching (SM) for further similarity calculation.

4. Experiments

In this section, we evaluate our method on a real-world Japanese and English dataset.

4.1 Experimental dataset

We employed metadata records of ukiyo-e prints in Japanese and English to validate our method.

We collected 203 ukiyo-e Japanese metadata records from Edo-Tokyo Museum [1] and 3,398 ukiyo-e English metadata records from the Metropolitan Museum of Art[2]. The metadata that were used in the experiments included artist names, titles, and series names of the ukiyo-e prints. In our dataset, every record has metadata of artist names and titles. A part of records has series names. Some examples of Japanese and English ukiyo-e metadata records are shown in Table 1 and Table 2, respectively.

In this dataset, each Japanese ukiyo-e metadata record has at least one corresponding English ukiyo-e metadata record in the English dataset, which means they refer to the same ukiyo-e print. For example, for the first Japanese metadata record in Table 1, its corresponding English metadata record is the first record in Table 2, since they refer to the same ukiyo-e print. To generate this ground truth data, for each Japanese ukiyo-e record, first, we utilized the ukiyo-e.org [3] image similarity analysis engine to find the most similar metadata records in the English dataset. Then, we manually checked whether the Japanese record and its most similar English record that is returned by ukiyo-e.org referred to the same ukiyo-e print.

Table 1: Some examples of Japanese ukiyo-e metadata records

作品名 (Title)	シリーズ名 (Series name)	作者 (Artist)
神奈川沖浪裏	冨嶽三十六景	葛飾北斎
深川万年橋下	冨嶽三十六景	葛飾北斎
日本橋 朝之景		歌川広重(初代)
隅田	雪月花	葛飾北斎
岡部 宇津之山	東海道五拾三次之内	歌川広重(初代)
女官洋服裁縫之図		橋本周延

4.2 Experimental setup

In the task of cross-language ukiyo-e prints linkage, we aimed to find record pairs that refer to the same ukiyo-e prints between datasets in Japanese and English.

Translation: In our experiments, we translated the titles and series names of ukiyo-e records from Japanese to English by using Microsoft Translator Text API[4]. As it provides two translation models: statistical machine translation (SMT) and neural network translation (NNT), we experimented with both translation models to translate metadata. The translation of metadata of our experimental dataset was made on April 13, 2018. Besides the metadata of the title and series name, we also utilized names of the artists of the ukiyo-e prints. Since artist names are not the target metadata of our proposed method, we translated the Japanese artist names by using a Japanese–English bilingual list of ukiyo-e artist names. This list was manually compiled using the authority data in the Web NDL Authorities[5], which is a web service provided by the National Diet Library (NDL), Japan.

Word embeddings: In our experiments, we utilized the skip-gram model of *word2vec* to learn word embeddings. To train the skip-gram model, we used the hyper-parameters recommended in [14], where the window size was 10 and the dimensionality of the word vectors was 200. The training data is the articles in English Wikipedia dump as of October 2017. The learned word embeddings contain 2,149,489 vocabularies.

[1] http://digitalmuseum.rekibun.or.jp/app/selected/edo-tokyo
[2] http://www.metmuseum.org/
[3] https://ukiyo-e.org/
[4] https://www.microsoft.com/en-us/translator/translatorapi.aspx
[5] http://id.ndl.go.jp/auth/ndla

Table 2: Some examples of English ukiyo-e metadata records

Title	Series name	Artist
Under the Wave off Kanagawa	Thirty-six Views of Mount Fuji	Katsushika Hokusai
Under the Mannen Bridge at Fukagawa	Thirty-six Views of Mount Fuji	Katsushika Hokusai
Morning View of Nihonbashi		Utagawa Hiroshige I
Snow on the Sumida River	Snow, Moon, and Flowers	Katsushika Hokusai
Utsu Hill at Okabe		Utagawa Hiroshige I
Court Ladies Sewing Western Clothing		Hashimoto Chikanobu

Record pair comparison: Generally, the similarity between two records is calculated by comparing several metadata similarities. In our experiments, the similarity between two ukiyo-e records (S_R) was determined by combining the title similarity (S_{title}), series name similarity (S_{series}), and artist name similarity (S_{artist}), which is defined in Equation (5).

$$S_R = S_{artist}(\alpha \cdot S_{title} + (1-\alpha) \cdot S_{series}) \quad (5)$$

Here, α is the weight of title similarity ($0 \le \alpha \le 1$). We set α = 0.75. The empirical tuning results will be presented in Section 4.3. S_{artist} uses exact string matching. It means S_{artist} is set as 1 if the translation of the Japanese artist's name is exactly the same with the English artist's name. Otherwise, S_{artist} is set as 0. Since the titles and series names are the target metadata of our method, S_{title} and S_{series} were calculated using our method or one of the following baseline methods.

Baseline methods: We compared our method (SS+SM) with the following two methods.

- **Soft-TFIDF:** This method [24] combines the TFIDF and Jaro-Winkler [12] measures. It first applies Jaro-Winkler (sim') to all pairs of words between two strings S and T, and then applies the TFIDF to words that have a similarity score above the threshold ($\theta \ge 0.9$) according to the Jaro-Winkler metric. Let $CLOSE(\theta, S, T)$ be the set of words $u \in S$ such that there is some $v \in T$ and $sim'(u, v) > \theta$, and for $u \in CLOSE(\theta, S, T)$, let $N(u,T) = max_{v \in T} sim'(u,v)$. The definition of Soft-TFIDF is shown in Equation (6).

$$Soft\text{-}TFIDF(S, T) = \sum_{u \in CLOSE(\theta,S,T)} wt(u,S) \cdot wt(v,T) \cdot N(u,T) \quad (6)$$

  where $wt(u,S)$ is the TFIDF weight of word u in S. We chose Soft-TFIDF as the baseline method to be compared with our method because it showed the best performance in title matching against 20 other commonly used string-based similarity measures [25].
- **SM:** This is the approach for semantic matching (SM) in our method. Here, we only used SM to calculate the similarity between translated metadata and metadata in target language without candidate identification.

Record pair classification: We consider cross-language record linkage as a ranking problem in our experiments. For each Japanese metadata record, we ranked candidate English metadata records according to the similarity score between them, which was calculated by Equation (5). Thus, we evaluated the ranking results in terms of Precison@n (P@n) and Recall@n (R@n).

4.3 Experimental results

Table 3 and Table 4 show the overall experimental results of using NNT translation models to translate metadata. We can see that our method performed better than the two baseline methods in terms of both P@n and R@n. It means that the methods that only use a string-based similarity metric or a semantic matching method are not sufficient for determining similar metadata. Our method achieved the highest P@1 of 57.64% and significantly outperformed others in terms of P@1 and R@1. It shows that our proposed method outperforms others when the top-ranked records are determined as the corresponding records in the target language.

Comparing with Soft-TFIDF, our method performs better. It indicates that our method that employs word embeddings can make up the insufficiency of string-based similarity metrics when measuring similarities between translated metadata and metadata in the target language.

Comparing our proposed method, SS+SM, with SM, it gains more than 10% P@1 and 9% R@1 improvement. It proves that our method using SS as a supplementary to identify candidate metadata has a positive impact on the later step, semantic matching.

Table 3: Experimental results: P@n

	P@1 (%)	P@2 (%)	P@3 (%)	P@4 (%)	P@5 (%)
Soft-TFIDF	43.35	29.56	23.32	18.47	15.57
SM	47.29	32.02	22.99	17.86	14.38
SS+SM	**57.64**	**36.95**	**25.94**	**19.83**	**16.06**

Table 4: Experimental results: R@n

	R@1 (%)	R@2 (%)	R@3 (%)	R@4 (%)	R@5 (%)
Soft-TFIDF	38.51	50.00	57.96	61.08	63.34
SM	43.68	53.04	57.14	59.11	59.61
SS+SM	**52.46**	**60.96**	**64.33**	**65.23**	**65.72**

In our experiments, one important parameter was the weight of title similarity α in the similarity metric of record pair comparison. Different weights of title similarity may lead to different performances in cross-language record linkage. Thus, we conducted experiments by varying the weight of title similarity. The bigger the value of α, the more weight given to title similarity and the less weight given to series name similarity. When $\alpha = 1$, it only considers title and artist name similarities in record pair comparison. Figure 5 and Figure 6 show the P@1 and R@1 performance of three methods with different values of title similarity weight α. The highest P@1 and R@1 are achieved by our proposed method on the title weight $\alpha = 0.75$. The performance improves by increasing α when $\alpha < 0.75$. Comparing $\alpha = 1$ and $\alpha = 0.75$, the performance of soft-TFIDF and our proposed method improves while SM drops. It indicates that the string-based similarity is a crucial part for calculating metadata similarity.

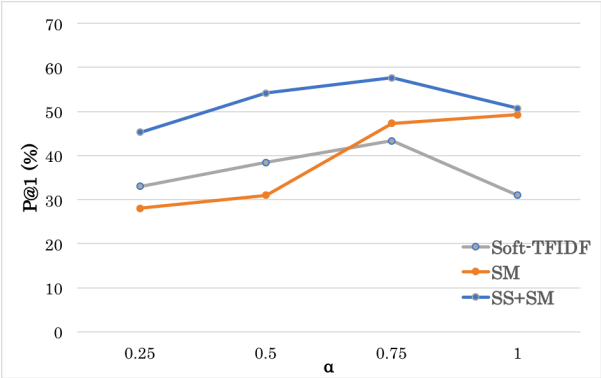

Figure 5: P@1 value vs. the weight of title similarity α

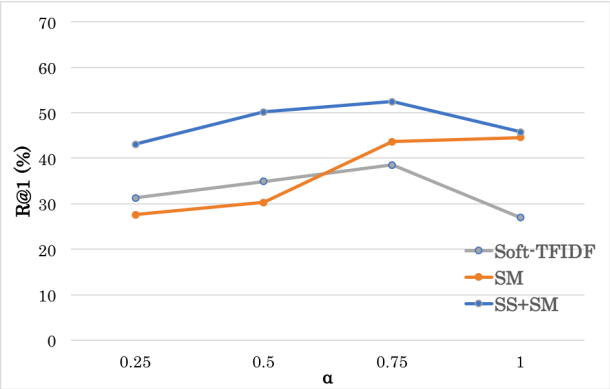

Figure 6: R@1 value vs. the weight of title similarity α

We also conducted experiments to study the effect of different translation models on cross-language record linkage. Table 5 shows the P@1 and R@1 performance of three methods with the translation models: STM and NNT. We can see that the results of using NNT are better than STM. It indicates that the performance of cross-language record linkage is influenced by the translation quality. Again, our method performs better when compared with two baseline methods. It shows that our method's performance is more stable; in other words, it is less affected by translation quality than others.

Table 5: Results of using two translation models

	P@1 (%)		R@1 (%)	
	STM	NNT	STM	NNT
Soft-TFIDF	36.45	43.35	34.07	38.51
SM	40.89	47.29	37.60	43.68
SS+SM	**56.16**	**57.64**	**51.89**	**52.46**

4.4 Case studies

We conducted some further case studies to analyze the effectiveness of our method.

Compared with Soft-TFIDF, our method performed better on the metadata with the words that are more likely to be translated with near-synonyms. In Figure 7, the word "図" in the Japanese title was translated into "figure" by using Bing Microsoft Translator. However, the corresponding word "図" is "scene" in the English title. The word "figure" and "scene" can be matched using our method because their meanings are similar. However, Soft-TFIDF fails because the word "figure" and "scene" are in different textual representations.

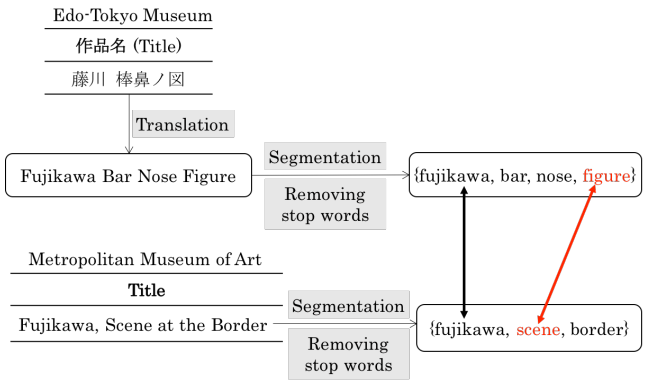

Figure 7: An example of how our method performs better than Soft-TFIDF

Compared with our method, SM's performance sometimes decreased since it tends to match semantically similar metadata. For example, given a record pair that refers to the same ukiyo-e print with the Japanese title "箱根 湖水図" and English title "Hakone Kosui", the Japanese title is translated into "Hakone Lake Map". The baseline method SM tends to match the title like "Sumida River in the Snow", since SM tries to match "Sumida" for "Hakone", and "River" for "Lake". In this case, our method can eliminate such English titles during candidate identification, since the translated title "Hakone Lake Map" and the title "Sumida River in the Snow" do not share the same words.

5. Related work

Our work on cross-language record linkage is related to cross-language entity linking [26][27] to some extent,

aiming to link the named entities in texts in one language to a knowledge base in another language. In this task, a lot of contextual information of named entities in texts and the content of articles in knowledge bases can be employed. Our work focuses on record linkage where only the metadata can be utilized, which is usually in short texts.

Cross-language knowledge linking [28][29] is another related task, which is creating links between articles in different languages that report on the same content. Most of the proposed methods use the structural information of data, such as inlink and outlink in the articles [28], to find the identical articles between knowledge bases in different languages. However, our approach aims to link the records across several databases in different languages that refer to the same real-world object, not to find the identical lexicons or articles.

Our work is also related to cross-language ontology matching. With the development of the Linked Data, ontology matching is attracting the interest of researchers. Cross-language ontology matching aims to find equivalent elements between two semantic data sources [16][30][31]. The difference between our goal and theirs is that our work focuses on general relational databases rather than semantic data sources.

6. Conclusions

We proposed a method of addressing the mismatching problem in cross-language record linkage. Our method specifically focuses on descriptive metadata such as titles and abstracts, which contain words that are sometimes translated into semantically similar words. To avoid unnecessary comparison of metadata, our proposed method first employs a simple but effective string-based similarity measurement to identify possible matched metadata as candidates. We then employ a semantic matching method to calculate the similarities between the translated metadata and their candidate metadata in the target language. Finally, for each translated metadata, its similar metadata in the target language are determined from these candidates by performing semantic matching. Our method makes up the insufficiency of string-based similarity metrics in measuring metadata similarity in cross-language record linkage. Through the experiments on a real-world dataset, we demonstrated that our method performs better than the baseline methods that only rely on string-based similarities or semantic matching.

In the future, we plan to improve our method by employing external knowledge resources when comparing the metadata in different languages. We also plan to validate the effectiveness of our method on the dataset in other languages.

[Acknowledgements]

This work was supported in part by JSPS KAKENHI Grant Number JP16K00452, and MEXT-Supported Program for the Strategic Research Foundation at Private Universities (S1511026).

[References]

[1] I. P. Fellegi and A. B. Sunter, “A Theory for Record Linkage,” *J. Am. Stat. Assoc.*, vol. 64, no. 328, pp. 1183–1210, Dec. 1969.

[2] N. Koudas, S. Sarawagi, and D. Srivastava, “Record linkage: Similarity Measures and Algorithms,” in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, 2006, pp. 802–803.

[3] S. Tejada, C. A. Knoblock, and S. Minton, “Learning domain-independent string transformation weights for high accuracy object identification,” in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[4] A. Elmagarmid, P. Ipeirotis, and V. Verykios, “Duplicate Record Detection: A Survey,” *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 1, pp. 1–16, 2007.

[5] B. Batjargal, T. Kuyama, F. Kimura, and A. Maeda, “Identifying the Same Records across Multiple Ukiyo-e Image Database Using Textual Data in Dif fferent Languages,” in *Proceedings of the 14th ACM/IEEE Joint Conference on Digital Libraries*, 2014, pp. 193–196.

[6] V. I. Levenshtein, “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals,” *Sov. Phys. Dokl.*, vol. 10, no. 8, pp. 707–710, 1966.

[7] F. J. Damerau, “A Technique for Computer Detection and Correction of Spelling Errors,” *Commun. ACM*, vol. 7, no. 3, pp. 171–176, 1964.

[8] P. Jokinen, J. Tarhio, and E. Ukkonen, “A Comparison of Approximate String Matching Algorithms,” *Softw. Pract. Exp.*, vol. 26, no. 12, pp. 1439–1458, 1996.

[9] K. Kukich, “Technique for Automatically Correcting words in Text,” *ACM Comput. Surv.*, vol. 24, no. 4, pp. 377–439, 1992.

[10] H. Keskustalo, A. Pirkola, K. Visala, E. Leppänen, and K. Järvelin, “Non-Adjacent Digrams Improve Matching of Cross-Lingual Spelling Variants,” in *Proceedings of the 10th International Symposium on String Processing and Information Retrieval*, 2003, pp. 252–265.

[11] M. A. Jaro, “Advances in Record Linkage Methodology as Applied to the 1985 Census of Tampa Florida,” *J. Am. Stat. Assoc.*, vol. 84, no. 406, pp. 414–420, 1989.

[12] W. E. Winkler, “String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage,” in *Proceedings of the Section on Survey Research, American Statistical Association*, 1990, pp. 354–359.

[13] A. E. Monge and C. P. Elkan, “The Field Matching Problem: Algorithms and Applications,” in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, no. Slaven 1992, pp. 267–270.

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

2013.

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

[16] B. Fu, R. Brennan, and D. O. Sullivan, "Cross-Lingual Ontology Mapping – An Investigation of the Impact of Machine Translation," in *Proceedings of the Asian Semantic Web Conference*, 2009, pp. 1–15.

[17] G.-A. Levow, D. W. Oard, and P. Resnik, "Dictionary-based Techniques for Cross-Language Information Retrieval," *Inf. Process. Manag.*, vol. 41, no. 3, pp. 523–547, 2005.

[18] A. Barrón-Cedeño, P. Gupta, and P. Rosso, "Methods for cross-language plagiarism detection," *Knowledge-Based Syst.*, vol. 50, pp. 211–217, 2013.

[19] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[20] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, 2011.

[21] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with compositional vector grammars," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, vol. 1, pp. 455–465.

[22] P. S. Dhillon, D. Foster, and L. Ungar, "Multi-view learning of word embeddings via CCA," in *Proceedings of Advances in Neural Information Processing System (NIPS 2011)*, 2011, pp. 1–9.

[23] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.

[24] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string metrics for matching names and records," in *Proceedings of the International Workshop on Data Cleaning and Object Consolidation, held at KDD*, 2003, pp. 73–78.

[25] N. Gali, R. Mariescu-Istodor, and P. Fränti, "Similarity Measures for Title Matching," in *Proceedings of 23rd International Conference on Pattern Recognition*, 2017, pp. 1549–1554.

[26] P. McNamee, J. Mayfield, D. Lawrie, D. W. Oard, and D. S. Doermann, "Cross-Language Entity Linking," in *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 2011, pp. 255–263.

[27] J. Mayfield, D. Lawrie, P. McNamee, and D. W. Oard, "Building a Cross-Language Entity Linking Collection in Twenty-One Languages," in *Proceedings of the Cross Language Evaluate Forum*, 2011, pp. 3–13.

[28] Z. Wang, J. Li, Z. Wang, and J. Tang, "Cross-lingual knowledge linking across wiki knowledge bases," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 459–468.

[29] R. Navigli and S. Ponzetto, "BabelNet: Building a very large multilingual semantic network," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 216–225.

[30] J. Li, J. Tang, Y. Li, and Q. Luo, "RiMOM: A dynamic multistrategy ontology alignment framework," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 8, pp. 1218–1232, 2009.

[31] J. Tang, J. Li, B. Liang, X. Huang, Y. Li, and K. Wang, "Using Bayesian decision for ontology mapping," *Web Semant.*, vol. 4, no. 4, pp. 243–262, 2006.

Yuting SONG

She is a Ph.D. candidate at Ritsumeikan University. She received a B.S. degree in Computer Science from Dalian University of Foreign Languages, China, in 2012 and M.S. degrees in Information Science from Beijing Language and Culture University, China, in 2015. Her research interests include natural language processing, semantic representation, and cross-language record linkage.

Biligsaikhan BATJARGAL

He is a Senior Researcher at the Kinugasa Research Organization, Ritsumeikan University. He received a Bachelor Degree in Computer Science from the Mongolian University of Science and Technology in 1999, received a Master Degree of Engineering in Advanced Information Science and Engineering from Ritsumeikan University in 2008, and received a Doctoral Degree in Engineering from Ritsumeikan University in 2012, respectively. His research interests include digital libraries, digital humanities, and multilingual information retrieval.

Akira MAEDA

He is a professor in the College of Information Science and Engineering, Ritsumeikan University. He received B.A. and M.A. degrees in Library and Information Science from the University of Library and Information Science in 1995 and 1997 and received the Ph.D. degree in Engineering from the Nara Institute of Science and Technology in 2000. His research interests include digital libraries, digital humanities, information retrieval, and multilingual information processing.