Metadata Similarity Calculation in Cross-Language Record Linkage based on Cross-lingual Embedding Models

Yuting SONG† Biligsaikhan BATJARGAL‡ and Akira MAEDA†‡

† Research Organization of Science and Technology, Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga, 525-8577 JAPAN
‡ Kinugasa Research Organization, Ritsumeikan University
56-1 Toji-in Kitamachi, Kita-ku, Kyoto, 603-8577 JAPAN
†‡ College of Information Science and Engineering, Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga, 525-8577 JAPAN

E-mail: † ytsong@gst.ritsumei.ac.jp, ‡ biligee@fc.ritsumei.ac.jp, † ‡ amaeda@is.ritsumei.ac.jp

Abstract In many data mining applications, there is a need to compare or combine information from different data sources. One of the main processes is to find identical records that refer to the same real-world entity across databases, which is known as record linkage or record matching. In this paper, we focus on the task of cross-language record linkage, in which records are from the databases in different languages. The descriptive metadata of an entity (e.g. the title of a book, the title of an image) summarize the content and distinguish it from other entities. Thus, we propose a method of measuring descriptive metadata similarities in different languages for record linkage. Our method utilizes cross-lingual embedding models, in which words and metadata in different languages can be represented in a shared embedding space. In this work, we learn the shared embedding space using sentence-level bilingual parallel data. We evaluate the effectiveness of our proposed method on the real-world databases in Japanese and English.

Keyword Cross-language record linkage, title matching, semantic similarities

1. Introduction

Record linkage [1][2][3] is a task of finding record pairs that refer to the same entities across multiple data sources, which is an important step in many data mining applications as information from multiple sources needs to be integrated or combined in order to allow more detailed data analysis. It can be used to improve data quality and to reduce costs and efforts in data acquisition [4].

With the World Wide Web becomes widely matured in more and more countries, the information is being produced in a variety of languages. The identical entities can exist in multiple data sources in different languages. An example is shown in Figure 1. The identical ukiyo-e prints[1] are digitized not only in the Japanese digital museums with the metadata in Japanese, but also in the digital museums of foreign countries with metadata in their native languages [5].

To find the record pairs that refer to the same entity, the record pairs are compared based on their metadata similarities. The task of cross-language record linkage is challenging since the metadata are in different languages.

In this paper, we focus on cross-language record linkage by measuring descriptive metadata similarities, since the descriptive metadata given to an entity summarizes and distinguishes it from other entities. Our method utilizes cross-lingual embedding models, in which words and metadata in different languages can be represented in a shared embedding space. Then, for calculating the metadata similarities across languages, we learn a linear mapping between vector spaces of languages to transform the metadata representations from the vector space of one language to the other.

The reminder of this paper is organized as follows. Section 2 outlines some related work; Section 3 introduces

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

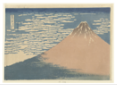
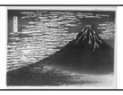
Ukiyo-e prints	Metadata		Language	Database
	Title	Artist		
	凱風快晴	葛飾北斎	Japanese	Edo-Tokyo Museum
	Gaifū kaisei	Katsushika Hokusai	English (Transliteration)	Library of Congress
	South Wind, Clear Sky	Katsushika Hokusai	English	Metropolitan Museum of Art
	Vent frais par matin clair	Hokusai Katsushika	French	French Photo Agency
	Helder weer en een zuidelijke wind	Katsushika Hokusai	Dutch	Rijksmuseum
	Fuji bei schönem Wetter von Süden gesehen	Katsushika Hokusai	German	Bildarchiv Foto Marburg

Figure 1: An example of the same ukiyo-e prints that are exhibited in multiple databases with metadata in different languages

our proposed method in detail; Section 4 presents our experimental setup and evaluations. Section 5 concludes this work and outlines future work.

2. Related work

2.1 Record linkage

Record linkage is the task of identifying records that refer to the same entities from several data sources. Over the past decade, various research fields have developed their own solutions to the problem of record linkage, and as a result, this task is named by many different terms. In the database field, record linkage and identity resolution [6] are used to describe the process of identifying the records that represent the same entities. When matching records are found, identity resolution merges the identical records, while record linkage simply notes the correspondence. In the natural language processing field, this problem is known by the name of coreference resolution [7][8], which is to determine two entity mentions refer to the same entity within document; and entity linking [9], which is to link the entities mentioned in text to the entry in a knowledge base.

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 the descriptive metadata, such as title.

2.2 Cross-language tasks

Cross-language entity linking [10][11] is related to our work to some extent, which aims to link the named entities in the texts in one language to a knowledge base in another language. In this task, much contextual information of named entities in texts and content of articles in knowledge bases can be employed. However, our work focuses on the record linkage where only the metadata values can be utilized, which are usually short texts, and sometimes in poor quality.

Cross-language knowledge linking [12][13] is another related task. Most methods are proposed using the structural information of data, such as inlink and outlink in the articles [12], to find the identical articles between knowledge bases in different languages. BabelNet [13] is a large multilingual lexical knowledge base built by combining Wikipedia and WordNet. However, our approach aims at linking the records in several databases in different languages that refer to the same real-world entity, not to find identical lexicons or articles.

Our work is also related to cross-language ontology matching [14][15][16]. With the development of the Linked Data[2], ontology matching is attracting the interests of some researchers. Cross-language ontology matching is to find equivalent elements between two semantic data

[2] http://linkeddata.org/

sources. The difference between our goal and theirs is that our work focuses on general relational databases.

3. Methodology

Our method utilizes cross-lingual embedding models, in which words and metadata in different languages can be represented in a shared embedding space.

In this section, first, we introduce two methods for the shared embedding space induction. Then, we explain in detail the metadata representations and metadata similarity calculation of our proposed method.

3.1 Cross-lingual embeddings models

Cross-lingual embeddings from bilingual word pairs. This type of models [17][18][19] focuses on learning the mappings between independently trained monolingual embedding spaces using a set of bilingual word pairs.

Cross-lingual embeddings from bilingual sentence parallel data. This type of models exploits some sentence-aligned parallel corpora to induce the cross-lingual embedding spaces. In our proposed method, we use the cross-lingual embeddings space that is induced by sentence-aligned parallel corpora.

3.2 Metadata similarity calculation

With the induced cross-lingual spaces we can directly measure the semantic similarity of words in two languages, but we still need to define how to represent metadata. To this end, we outline the method that exploits the induced cross-lingual embedding space for metadata similarity calculation.

We represent metadata as vectors by adding the cross-lingual embeddings of their constituent words. We opt for vector addition as composition since word embedding spaces exhibit linear linguistic regularities [20]. Thus, our method for representing metadata can be formulated in the following equation:

$$R(M) = \frac{1}{n}\sum_{i=1}^{n} w_i \quad (1)$$

where n is the number of words in metadata M; w_i is the vector embeddings of words that compose the metadata M.

An example in Figure 2 illustrates the process of metadata representation. First, the vector embeddings of words in the English title “Storm below Mount Fuji” are obtained from the pre-trained English word embeddings. Next, the metadata is represented as a vector by adding the obtained vector embeddings.

We use cross-lingual embeddings to transform the metadata vectors from one language vector space to the vector space of the other language. We learn a linear mapping between the vector spaces in different languages using bilingual sentence pairs. Suppose we have a set of bilingual sentence pairs and their associated vector representations $\{x_i, y_i\}_{i=1}^{n}$, where x_i is the vector of sentence i in the source language, and y_i is the vector of its corresponding sentence in the target language. Our goal is to learn a mapping matrix W such that Wx_i approximates y_i.

At the time of similarity calculation, for any given new metadata vector x, we transform it into the vector space of the other language by computing $z = Wx$. Then, we can calculate the similarity between metadata in different languages by comparing the transformed vectors with other metadata vectors in the vector space of the other language.

4. Experiments

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

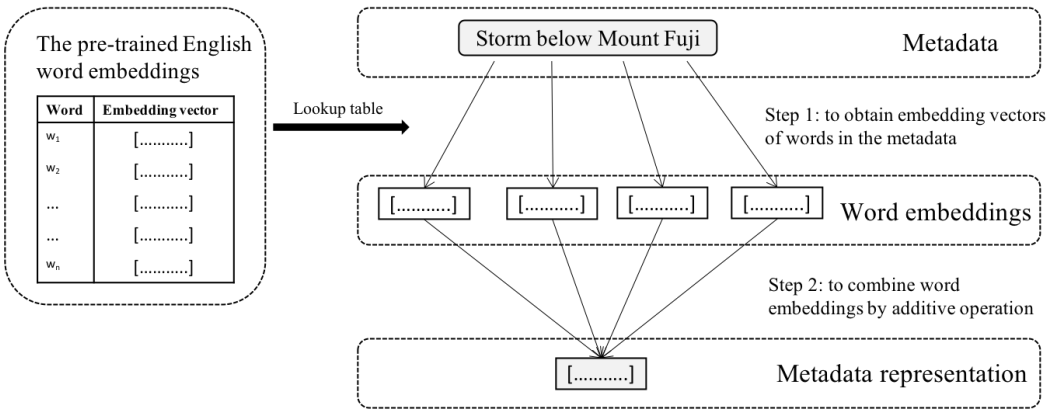


Figure 2: An example of metadata representation

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

作品(Title)	シリーズ名(Series name)	作者(Artist)	制作年(Date)
神奈川沖浪裏	冨嶽三十六景	葛飾北斎	天保 2 年～4 年
深川万年橋下	冨嶽三十六景	葛飾北斎	天保 2 年～4 年
女官洋服裁縫之図		橋本周延	明治 20 年 8 月
日本橋 朝之景		歌川広重	天保中期
木場の雪		歌川国貞	文化 8 年～天保末
隅田	雪月花	葛飾北斎	天保 3 年

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

Title	Series name	Artist	Date
Under the Wave off Kanagawa	Thirty-six Views of Mount Fuji	Katsushika Hokusai	ca. 1830–32
Morning View of Nihonbashi		Utagawa Hiroshige I	ca. 1833–34
Court Ladies Sewing Western Clothing		Hashimoto Chikanobu	August 23rd, 1887
Snow on the Sumida River	Snow, Moon, and Flowers	Katsushika Hokusai	ca. 1833
Utsu Hill at Okabe		Utagawa Hiroshige I	1834
Evening Glow at Koganei Border		Ryūryūkyo Shinsai	1797–1858

4.1 Experimental dataset

We collected 203 Japanese ukiyo-e metadata records from Edo-Tokyo Museum[3] and 3,398 English ukiyo-e metadata records from the Metropolitan Museum of Art[4]. The metadata that was used in the experiments includes artist names, titles, series names and date of the ukiyo-e prints. In our dataset, every record has metadata of artist names and titles. Some metadata records have series names and date. 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 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 identified by ukiyo-e.org[5] referred to the same ukiyo-e print.

4.2 Experimental setup

4.2.1 Word embeddings of different languages

In the experiments, both English word embeddings and Japanese word embeddings are trained using Word2vec toolkit. The skip-gram model of Word2vec is employed to learn word embeddings. To train the skip-gram model, the hyper-parameters recommended in [21] are used, which are shown in Table 3.

Table 3: The parameters of Word2vec for training word embeddings

Model	Skip-gram
Window size	10
Vector dimensionality	200

To train English word embeddings, English Wikipedia articles are used, which are the data in English Wikipedia dump[6] as of September 2018.

To train Japanese word embeddings, Japanese Wikipedia articles are used, which are the data in Japanese Wikipedia dump as of September 2018.

4.2.2 Parallel sentence corpora

To learn the mapping between the vector spaces that

[3] http://digitalmuseum.rekibun.or.jp/app/selected/edo-tokyo

[4] http://www.metmuseum.org/

[5] https://ukiyo-e.org/

[6] https://dumps.wikimedia.org/

represent Japanese and English, 600 Japanese-English parallel short sentence pairs are used. These Japanese-English parallel short sentence pairs are extracted from Tanaka corpus[7]. The lengths of these parallel short sentence pairs range from 6 to 12 words, which are equivalent to the length of ukiyo-e titles. Some examples of Japanese-English parallel short sentence pairs are shown in Table 4.

Table 4: Some examples of Japanese-English parallel short sentence pairs

Japanese sentence	**English sentence**
私はテニス部員です	I'm in the tennis club
多くの動物が人間によって滅ぼされた	Many animals have been destroyed by men
私達は国際人になりたいと思います	We want to be international
彼の小説は１つも読んでいない	I haven't read any of his novels
私は音楽が好きではありません	I do not like music
申告する物は何もありません	I have nothing to declare
父は今家にいるだろう	My father may be at home now
近くで火事が起こった	A fire broke out nearby
スケートの方が好きです	I like skating better
学生全員が出席した	All of the students were present

In order to make the learned mapping more accurate to transform the metadata of ukiyo-e prints, the experiments further use several pairs of Japanese and English ukiyo-e titles to optimize the learned mapping, in which each Japanese and English title pair refers to the same ukiyo-e prints.

4.2.3 Baseline methods

We compare our proposed method with the performance of cross-language record linkage with the translation-based method. In the translation-based method, we translated the titles and series names of ukiyo-e records from Japanese to English by using Microsoft Translator Text API[8]. 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 January 22, 2019.

4.2.3 Record pair comparison

Besides the metadata of the title and series name, we also utilized the artist name and date of the ukiyo-e prints.

Artist names. 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[9], which is a web service provided by the National Diet Library (NDL) of Japan.

Date. Since the dates metadata of ukiyo-e prints are numeric values, they are also not the target metadata of our proposed method. In the Japanese ukiyo-e datasets, the dates are represented in the Japanese calendar, such as the examples that are shown in Table 1. However, in the English ukiyo-e datasets, the dates are represented in the western calendar, such as the examples that are shown in Table 2. To compare the dates between the Japanese datasets and English datasets, we use HuTime Web API[10] to convert the dates in Japanese calendar to western calendar.

In the task of record linkage, the similarity between two records is calculated by comparing several metadata similarities. Here, the similarity between two ukiyo-e records (S_R) is determined by combining the title similarity (S_{title}), series name similarity (S_{series}), artist name similarity (S_{artist}) and date similarity (S_{date}), which is defined in the Equation (2).

$$S_R = S_{artist}(\alpha \cdot S_{title} + \beta \cdot S_{series} + \gamma \cdot S_{date}) \qquad (2)$$

Here, $\alpha + \beta + \gamma = 1$. α is the weight of title similarity; β is the weight of series name similarity; γ is the weight of date similarity.

S_{artist} uses the exact string matching. It means S_{artist} is set as 1 if the translation of the Japanese artist name is the same with the English artist name. Otherwise, S_{artist} is set as 0.

S_{date} is defined as follows:

[7] Tanaka corpus consist of Japanese-English parallel sentence pairs, which are collected from Japanese−English bilingual newspaper articles and broadcast media news reports published on the WWW.
http://www.edrdg.org/wiki/index.php/Tanaka_Corpus

[8] https://www.microsoft.com/en-us/translator/translatorapi.aspx

[9] http://id.ndl.go.jp/auth/ndla

[10] http://ap.hutime.org/cal/index.html

$$S_{date} = \begin{cases} 1 - \frac{|d_j - d_e|}{d_{max}}, & if\ |d_j - d_e| < d_{max} \\ 0, & else \end{cases} \quad (3)$$

Here, d_j is Japanese date; d_e is English date.

Since title and series name are the target metadata of this work, S_{title} and S_{series} are calculated using our proposed method that is introduced in Section 3. In the comparison experiments, S_{title} and S_{series} are calculated using the baseline methods that are introduced in Section 4.2.3.

4.2.4 Evaluation

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. Thus, we evaluated the ranking results in terms of Precison@n (P@n) and Recall@n (R@n).

4.3 Experimental results

The experimental results are shown in Table 5 and Table 6.

Table 5: The experimental results: P@n

	P@1(%)	P@2(%)	P@3(%)	P@4(%)	P@5(%)
Translation-based method (STM)	51.72	35.71	25.12	19.46	16.16
Translation-based method (NNT)	56.15	35.47	26.11	20.32	16.45
Our proposed method	55.29	34.12	21.54	18.22	14.53

Table 6: The experimental results: R@n

	R@1(%)	R@2(%)	R@3(%)	R@4(%)	R@5(%)
Translation-based method (STM)	47.62	61.66	65.35	66.75	68.88
Translation-based method (NNT)	51.72	61.01	66.17	67.89	68.88
Our proposed method	51.19	52.54	54.19	58.32	61.22

Comparing two baseline methods, it can be seen that the results of using NNT are better than STM. It indicates that the performance of cross-language record linkage is affected by the translation quality.

Although the P@1 and R@1 of our proposed method are a bit lower than the translation-based method (NNT), the method without translation has less bilingual data requirements, which could possibly be applied to other low-resourced languages.

5. Conclusion

In this paper, we presented a method to measure the similarity between metadata in different languages for cross-language record linkage, which does not use any translation methods. This method only uses a small set of bilingual parallel data to learn a linear mapping between the vector space of the source language and vector space of the target language, which is used to transform the vector representations of metadata from the source language to the target language.

In the future, we plan to apply our method to other datasets, such as book and film datasets. We also plan to validate the effectiveness of our method on the dataset in other languages.

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] S. Sarawagi and A. Bhamidipaty, “Interactive deduplication using active learning,” in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD 2002)*, 2002, pp. 269–278.

[3] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg, “Adaptive name matching in information integration,” *IEEE Intell. Syst.*, vol. 18, no. 5, pp. 16–23, 2003.

[4] P. Christen, “A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 9, pp. 1537–1555, 2012.

[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] S. Bartunov, A. Korshunov, S. Park, W. Ryu, and H. Lee, “Joint Link-Attribute User Identity Resolution in Online Social Networks Categories and Subject Descriptors,” in *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis*, 2012.

[7] R. Collobert and J. Weston, “Deep Learning for Natural Language Processing,” *Slides*, pp. 1–113, 2009.

[8] V. Ng and C. Cardie, “Improving Machine Learning Approaches to Coreference Resolution,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 104–111.

[9] A. Moro, A. Raganato, and R. Navigli, “Entity Linking meets Word Sense Disambiguation: a Unified Approach,” *Trans. Assoc. Comput. Linguist.*, vol. 2, pp. 231–244, 2014.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] B. Fu, R. Brennan, and D. O’sullivan, “Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web.,” in *In Proceedings of WWW Workshop on Multilingual Semantic Web*, 2010, pp. 13–20.

[15] 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.

[16] 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.

[17] M. Artetxe, G. Labaka, and E. Agirre, “Learning bilingual word embeddings with (almost) no bilingual data,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 451–462.

[18] T. Mikolov, Q. V Le, and I. Sutskever, “Exploiting Similarities among Languages for Machine Translation,” in *arXiv preprint arXiv:1309.4168v1*, 2013, pp. 1–10.

[19] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla, “Offline bilingual word vectors, orthogonal transformations and the inverted softmax,” in *In Proceedings of the International Conference for Learning Representations 2017*, 2017, pp. 1–10.

[20] T. Mikolov, W.-T. Yih, and G. Zweig, “Linguistic Regularities in Continuous Space Word Representations,” in *Proceedings ofNAACL-HLT 2013*, 2013, pp. 746–751.

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