Recognition and Transliteration of Proper Nouns in Cross-Language Record Linkage by Constructing Transliterated Word Pairs

Yuting Song, Biligsaikhan Batjargal and Akira Maeda
Graduate School of Information Science and Engineering, Ritsumeikan University
Research Organization of Science and Technology, Ritsumeikan University
College of Information Science and Engineering, Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga, 525-8577, Japan
gr0260ff@ed.ritsumei.ac.jp, biligee@fc.ritsumei.ac.jp, amaeda@is.ritsumei.ac.jp

Abstract

Proper nouns in metadata are representative features for linking the identical records across data sources in different languages. To improve the recognition of proper nouns in metadata and obtain their transliterations, we propose a method to construct bilingual transliteration word pairs, in which transliterated words in target language are back-transliterated to their original words in source language. The acquired transliterated word pairs are employed to recognize and transliterate proper nouns in metadata. We evaluated our proposed method on the task of cross-language record linkage between a Japanese database and an English database. Experimental results show the usage of the transliterated word pairs that we have obtained can improve the effectiveness of cross-language record linkage.

Keywords

Proper noun recognition, Back-transliteration, Cross-language record linkage.

1. Introduction

In many knowledge discovery and data mining applications there is a need to combine information from heterogeneous sources. The core issue is to find the identical records that refer to the same real-world entity across multiple sources, which is known as record linkage or record matching (Fellegi and Sunter 1969; Sarawagi and Bhamidipaty 2002; Bilenko et al.

2003; Elmagarmid et al. 2007). Within the context of globalization, the multilingual data sources make record linkage process more challenging, since identical records could be from the sources in different languages.

Comparing the record pairs in different languages requires that the metadata of records in one language be translated into another language. Proper nouns in metadata are important for record comparison, as they carry the distinctive information in a metadata. Thus, accurate recognition and transliteration of proper nouns are key components of cross language record linkage.

State-of-the-art Named Entity Recognition (NER) systems do not perform well on recognizing the proper nouns in metadata, since the metadata like titles, are usually very short and consist of a few words, particularly when metadata are describing the records from a specific domain. For example, using MeCab[1] to recognize the proper nouns in the title of a Japanese artwork "深川万年橋下", the results are shown in Table 1.

	品詞 (Part of speech)		
	大分類 (Major classification)	中分類 (Medium classification)	小分類 (Small classification)
深川	名詞 (Noun)	固有名詞 (Proper noun)	地名 (Place name)
万	名詞 (Noun)	数詞 (Number)	
年	名詞 (Noun)	普通名詞 (Common noun)	助数詞可能 (Possible counter word)
橋下	名詞 (Noun)	普通名詞 (Common noun)	一般 (General)

Table 1: The proper noun recognition results of a Japanese title

In this example, the proper noun "万年", which is the name of a bridge "万年橋" (Mannen bridge), is improperly segmented as two morphemes and mistaken as two common words.

In this article, we propose a method to recognize and transliterate the proper nouns in metadata of records, particularly the descriptive metadata such as title and abstract, which aims at improving the effectiveness of cross-language record linkage. Our work focuses on record linkage between Japanese databases and English databases, in which Japanese is the source language and English is the target language.

[1] MeCab is a Japanese part-of-speech and morphological analyzer.

Our proposed method is motivated by an observation that the English translation of a Japanese proper noun is mostly a transliterated word. For example, the English correspondence of a Japanese proper noun “万年” shown above is “Mannen”, which is a transliterated word. Based on this observation, we firstly identify the transliterated words in the metadata from the database in English. Then, we convert these transliterated words into their corresponding original Japanese words, which is a back-transliteration process. In this way, we could obtain pairs of Japanese words and their transliterated words, which are employed to recognize and transliterate the proper nouns in Japanese metadata.

This work is an extended version of our previous work (Song et al. 2016). We improve the previous methods by using a more reliable approach to identify the transliteration words in English metadata. We evaluate the effectiveness of our proposed method on a new dataset, which is closer to real world scenarios. Besides, we compare our proposed method against the baseline method with different parameters.

The rest of this article is organized as follows. In Section 2, related work is reviewed. Section 3 introduces the general process of cross-language record linkage. In Section 4, we describe our proposed method in detail. We evaluate our method through experiments in Section 5. Section 6 concludes our work.

2. Related Work

Proper noun recognition is similar to NER, which has been an important topic of study in natural language processing. Similar to the idea of our work, some research articles (Huang and Voge 2002; Li et al. 2012) have explored that the corpora in different languages could be used as complementary features to improve recognition performance. These methods require bilingual parallel corpora, which are costly to construct, or difficult to obtain. Comparing to them, our method is simpler since we only use the metadata in target language, which can be obtained easily in cross-language record linkage.

Some research has been conducted for the back-transliteration (Durrani et al. 2014), such as loanwords in Korean (Jeong et al. 1999; Kang and Choi 2000) and Japanese to English (Knight and Graehl 1998; Bilac and Tanaka 2004). Most of them have focused on training a statistical transliteration model using bilingual word lists, and the larger corpora have showed the better performance. However, in our method, we only use pairs of phrases and their pronunciations, which can be easily obtained from the encyclopedias in the related domain of records in the source language.

Cross-language entity linking (McNamee et al. 2011; Mayfield et al. 2011) is related to our work to some extent, which aims to link the same entities in one language to a knowledge base in another language. In this task, much contextual information of named entities and content of documents in a knowledge base can be employed. However, our work focuses on the record linkage where only the metadata can be utilized, which are short texts, and sometimes in poor quality.

3. Cross-language record linkage

In this section, we explain the cross-language record linkage process. Generally, record linkage identifies the records from disparate data sources that refer to the same entity by comparing several metadata of records. Different from monolingual record linkage, cross-language record linkage requires translating metadata in one language to another language in order to compare the records within the same language.

The general procedure of cross-language record linkage is shown in Figure 1. Given two databases in different languages (assuming one is source language, the other is target language), firstly, the metadata of records in source language are translated into the target language. Secondly, record pairs are compared by calculating the similarities between metadata. Finally, based on these metadata similarities, the record pairs are classified into matches and non-matches by using a certain classification model. The matched record pairs are regarded as the identical records that refer to the same real-world entities.

Our focus in this work is to recognize proper nouns in metadata in source language and obtain their appropriate transliterations, in order to improve the effectiveness of cross-language record linkage.

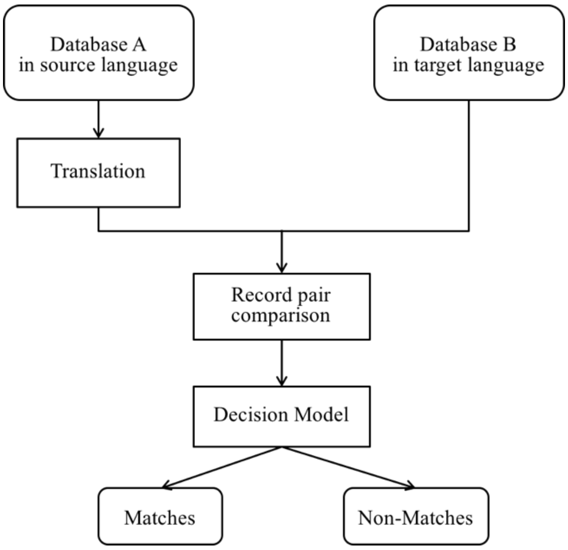


Figure 1: The general procedure of cross-language record linkage

4. Proper noun recognition and transliteration

We recognize proper nouns in metadata in source language and obtain their appropriate transliterations by constructing transliterated word pairs from metadata in target language. Figure 2 shows the process of our proposed method.

As we introduced in Section 1, our proposed method is motivated by an observation that the English translation of a Japanese proper noun is mostly a transliterated word. Based on this observation, we first identify the transliterated words from the metadata in English by using the spelling rules of Japanese words, which is introduced in Section 4.1. Then, in Section 4.2, we introduce the way to find the original Japanese words of transliterated words. Finally, we filter out some incorrect words from the acquired transliterated word pairs by adding the constraint on word composition, which will be explained later in Section 4.3.

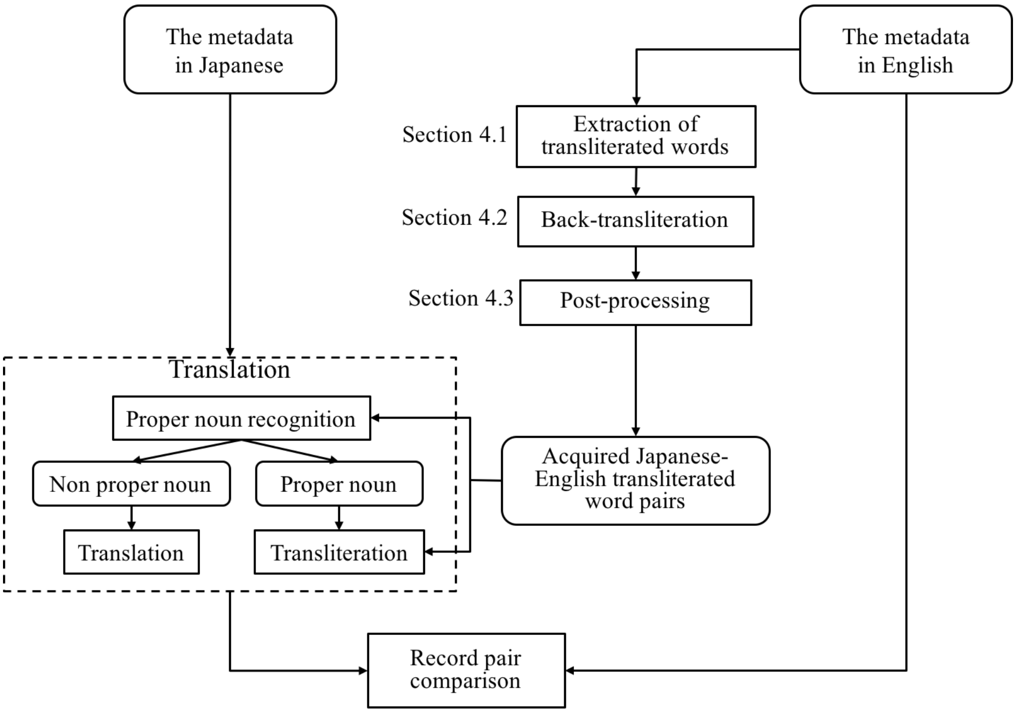


Figure 2: The process of our proposed method of proper noun recognition and transliteration in cross-language record linkage

4.1. Extraction of transliterated words

In this step, our goal is to extract transliterated words from English metadata. The transliterated words are the words that are romanized from Japanese words based on their pronunciations, such as “Yoshino” and “Fukagawa”. We first tokenize English metadata into words and convert uppercase to lowercase by using the Stanford Tokenizer (Manning et al. 2014). Then, we remove stop words via a stop word list, which includes function words and punctuations.

Since the pronunciation of a Japanese word consists of one or more Japanese syllables, we distinguish the transliterated word from English word (non-transliterated word) by judging whether it can be segmented by the Japanese syllabary, which is shown in Table 2. For example, the word “taiko” can be segmented into “ta”, “i” and “ko” using the Japanese syllabary, therefore, we judge it as a transliterated word. The word “good” cannot be segmented, since “d” cannot be found in the Japanese syllabary, we judge it as a non-transliterated word.

a	chi	mu	gu	pa	nyo	bya
i	tsu	me	ge	pi	hya	byu
u	te	mo	go	pu	hyu	byo
e	to	ya	za	pe	hyo	pya
o	na	yu	ji	po	mya	pyu
ka	ni	yo	zu	kya	myu	pyo
ki	nu	ra	ze	kyu	myo	
ku	ne	ri	zo	kyo	rya	
ke	no	ru	da	sha	ryu	
ko	ha	re	de	shu	ryo	
sa	hi	ro	dp	sho	gya	
shi	fu	wa	ba	cha	gyu	
su	he	o	bi	chu	gyo	
se	ho	n	bu	cho	ja	
so	ma	ga	be	nya	ju	
ta	mi	gi	bo	nyu	jo	

Table 2: Japanese syllabary (Hepburn Romanization)

4.2. Back-transliteration

After obtaining a set of transliterated words from metadata in English in the previous step, we convert these transliterated words back into their original Japanese words, which is a back-transliteration process.

We find the corresponding Japanese original words of transliterated words by using some Japanese phrases and their pronunciations pairs, which can be obtained from the encyclopedias of the related domain of Japanese records. This process is shown in Figure 3.

We represent the pronunciations of Japanese phrases in hiragana, which is one component of Japanese kana syllabary that can be used to represent pronunciations of Japanese words. In order to match these pronunciations that are represented in hiragana, English transliterated words are converted to corresponding Japanese hiragana sequences by using Hepburn Romanization system.

At the same time, we align the Japanese phrases with their pronunciations character by character. The characters here are Japanese kanji, which are logographic Chinese characters that are used in the Japanese writing system alongside hiragana and katakana. In this alignment process, the pronunciations of Japanese kanji are obtained from a Japanese kanji pronunciation dictionary. The corresponding pronunciation of each kanji in phrases is determined by shortest string matching in forward direction. Then, the transliterated words that are represented in hiragana are used to exactly or partially match the pronunciations of phrases. The corresponding Japanese words of exactly or partially matched pronunciations can be obtained according to the alignment results. These matched Japanese words are considered as the original words of transliterated words.

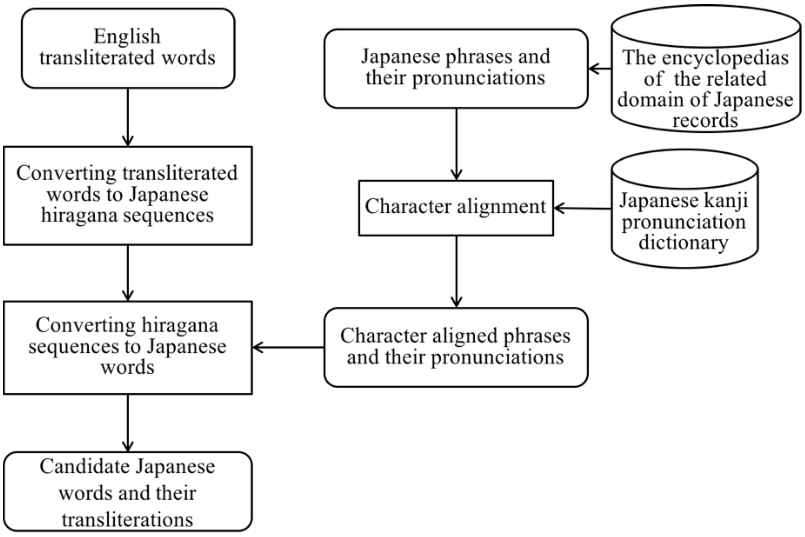

English transliterated words
Converting transliterated words to Japanese hiragana sequences
Converting hiragana sequences to Japanese words
Candidate Japanese words and their transliterations
Japanese phrases and their pronunciations
The encyclopedias of the related domain of Japanese records
Character alignment
Japanese kanji pronunciation dictionary
Character aligned phrases and their pronunciations

Figure 3: The back-transliteration process

To illustrate the process of back-transliteration, some examples are shown in Figure 4. The transliterated word “Sumida” is converted into Japanese hiragana “すみだ” by using Hepburn Romanization system. Then, “すみだ” is used to partially match the pronunciation “すみだがわ” of the phrase “隅田川”. Its corresponding Japanese word “隅田” is obtained according to the alignment result “隅 ｜ 田 ｜ 川 –– すみ ｜ だ ｜ がわ” . The Japanese word “隅田” is the original word of transliterated word “Sumida”.

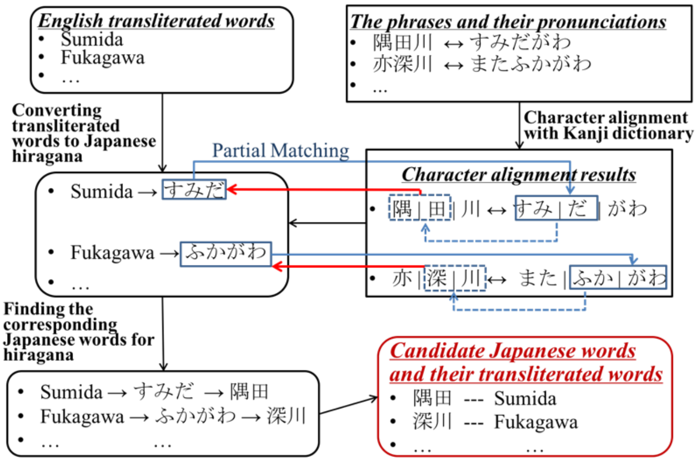

English transliterated words
Sumida
Fukagawa
…
The phrases and their pronunciations
隅田川 ↔ すみだがわ
亦深川 ↔ またふかがわ
…
Character alignment with Kanji dictionary
Converting transliterated words to Japanese hiragana
Partial Matching
Character alignment results
隅 | 田 | 川 ↔ すみ | だ | がわ
亦 | 深 | 川 ↔ また | ふか | がわ
Sumida → すみだ
Fukagawa → ふかがわ
…
Finding the corresponding Japanese words for hiragana
Sumida → すみだ → 隅田
Fukagawa → ふかがわ → 深川
… …
Candidate Japanese words and their transliterated words
隅田 --- Sumida
深川 --- Fukagawa
… …

Figure 4: Examples of back-transliteration

4.3. Post-processing

In the back-transliteration process, some transliterated words might be converted into incorrect proper nouns. For example, transliterated word “yoshino” can be back-transliterated into “吉野” (a place name), “葭の” or “吉の”. Among them, “葭の” and “吉の” are incorrect proper nouns. Such words can be partly eliminated by adding some constraints on word composition. We calculate the Japanese character distributions of proper nouns composition from a Japanese proper noun list. The majority of Japanese proper nouns mainly consist of kanji. Based on this characteristic of word composition of Japanese proper nouns, incorrect Japanese proper nouns could be filtered out by the position of kana, which is one component of the Japanese writing system along with kanji. If a Japanese kana appears at the end of a word, that word might be an incorrect proper noun. For example, “葭の” and “吉の” are determined as incorrect proper nouns since the Japanese kana “の” appears at the end of the words.

5. Experiments

In this section, we describe the experiments to evaluate the effectiveness of the proposed method for the task of cross-language record linkage.

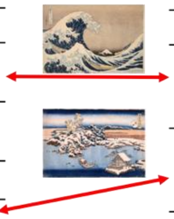
5.1. Data preparation

In order to evaluate our proposed method, we construct a dataset that contains metadata of Ukiyo-e print records from the websites of Edo-Tokyo Museum [2] (Japanese) and Metropolitan Museum of Art[3] (English).

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). These prints have been digitized and exhibited on the websites of many digital libraries and museums with metadata in various languages (Batjargal et al. 2014). Among these digital libraries and museums, some of record pairs refer to the same Ukiyo-e prints but their metadata are in different languages. Figure 5 shows some examples of same Ukiyo-e print records between a Japanese database and an English database.

A Japanese database

作品名	作者
冨嶽三十六景 神奈川沖浪裏	葛飾北斎
冨嶽三十六景 深川万年橋下	葛飾北斎
日本橋　朝之景	歌川広重(初代)
雪月花　隅田	葛飾北斎



An English database

Title	Artist
Under the Wave off Kanagawa, from the series Thirty-six Views of Mount Fuji	Katsushika Hokusai
Snow on the Sumida River, from the series, Snow, Moon, and Flowers	Katsushika Hokusai
Morning View of Nihonbashi	Utagawa Hiroshige

Figure 5: Examples of same Ukiyo-e print records between a Japanese database and an English database

The experimental dataset consists of 2555 Japanese titles and artist names of Ukiyo-e prints' records and 3408 English titles and artist names. Between the Japanese dataset and English dataset, there are 257 record pairs that refer to the same Ukiyo-e prints.

As we introduced in Section 4.2, the back-transliteration in our method requires domain related Japanese phrases and their pronunciations pairs. Since our experimental dataset is related to Ukiyo-e prints, we obtain Japanese phrases and their pronunciations pairs from five Ukiyo-e related encyclopedias in Japanese, including 《浮世絵百科》, 《浮世絵大事典》,

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

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

《浮世絵事典》[4], 《浮世絵鑑賞基礎知識》[5] and 《浮世絵百科(画題)》[6]. Besides, the phrases and their pronunciations are aligned by using the Japanese kanji dictionary[7].

5.2. Experimental setup

5.2.1 Translation

Here we translate the Japanese titles and artist names of Ukiyo-e records into English in order to compare with the records from English dataset.

- **Our proposed method**: we use our proposed method that are introduced in Section 4 to recognize and transliterate the proper nouns in Japanese titles.
- **Baseline method**: we also use MeCab (Kudo et al. 2004) to recognize proper nouns in Japanese titles as a baseline method. MeCab is a conditional random fields (CRFs) based Japanese part-of-speech and morphological analyzer. By using MeCab, the Japanese titles can be segmented into a set of words with the part-of-speech tags and pronunciations. In the baseline method, the recognized proper nouns are transliterated based on their pronunciations.

In both our method and baseline method, for the non-proper nouns in Japanese titles, we translate them by using EDR Japanese-English bilingual dictionary[8]. For translating Japanese artist names, we use BabelNet[9], which is a multilingual encyclopedic dictionary. By using BabelNet, we could obtain a set of corresponding English representations for each Ukiyo-e artist, including the transliterations and alias.

5.2.2 Record pairs comparison

After translating Japanese metadata to English, the similarity between two records is calculated by comparing several metadata. Here, we compare the titles and artist names.

- **Artist name comparison**: for comparing the artist names of two records, if the English artist name is included in the set of the Japanese artist name’s English representations, we set the similarity score of their artist name comparison as 1, otherwise the similarity score as 0.

[4] The encyclopedias 《浮世絵百科》, 《浮世絵大事典》, 《浮世絵事典》 contain the titles, artists, subjects, publishers and some other information of Ukiyo-e prints.

[5] 《浮世絵鑑賞基礎知識》 is an encyclopedia of the basic knowledge for appreciating Ukiyo-e.

[6] 《浮世絵百科(画題)》 is an encyclopedia of the titles of Ukiyo-e prints.

[7] http://www.kanjijiten.net/index.html

[8] http://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html

[9] http://babelnet.org/about

- **Title comparison**: the similarity calculation between the translation of Japanese title and English title is formulated as

$$\text{Similarity metric } = \frac{w_p \cdot M_p + w_{np} \cdot M_{np}}{w_p \cdot N_p + w_{np} \cdot N_{np}} \qquad (1)$$

  M_p, M_{np} are the number of matched proper nouns and non-proper nouns in a Japanese title respectively. w_p and w_{np} are their weights. N_p, N_{np} are the total number of proper nouns and non-proper nouns in a Japanese title respectively.

In our experiments, in order to reduce the complexity of record comparisons, we first compare the artist names of two records. If the similarity score of artist name comparison is 1, we further compare the titles. Otherwise, we give the similarity score of title comparison to 0.

5.2.3 Classification

Based on the similarity scores obtained in the previous step (Section 5.2.2), the record pairs are classified into matches and non-matches depending on the classification model used. The match here means that two records refer to the same real-world entity. The general idea is that larger similarity score between two records is, the more likely they refer to the same real-world entity. Here, we use a simple classification method, which applies a similarity threshold *t* to decide whether two records are matches or non-matches. If the similarity score between two records is larger than *t*, these two records are classified into matches. Otherwise, these two records are classified into non-matches.

5.2.4 Evaluation

The experimental results of cross-language record linkage are evaluated using the precision and recall (Christen et al., 2007). The definition of precision and recall are defined as

$$\text{Precision} = \frac{True\ matches}{True\ matches + False\ matches} \qquad (2)$$

$$\text{Recall } = \frac{True\ matches}{True\ matches + False\ non\text{-}matches} \qquad (3)$$

True matches is the number of record pairs that have been classified as matches and that are actual true matches. *False matches* is the number of record pairs that have been classified as matches, but they are not true matches. Actually, these record pairs refer to two different entities. The classifier has made a wrong decision with these record pairs. *False non-matches* is the number of record pairs that have been classified as non-matches, but they are actual true matches.

Precision calculates the proportion of how many of the classified matches (*True matches+False matches*) have been correctly classified as true matches (*True matches*). Thus, it measures how precise a classifier is in classifying true matches.

Recall calculates the proportion of true matches (*True matches+False non-matches*) that have been classified correctly (*True matches*). Thus, it measures how many of the actual true matching record pairs have been correctly classified as matches.

5.3. Experimental results

We conducted the experiments of identifying the same Ukiyo-e records between Japanese and English databases by using our proposed method and the baseline method. Table 3 shows the precision and recall of the experimental results when the similarity threshold $t = 0.6$. Here we set the weight of proper nouns $w_p = 2$ and the weight of non-proper nouns $w_{np} = 1$, which are decided experimentally. We observed that the experimental results are better when giving a larger weight to proper nouns than non-proper nouns. For example, in our experimental dataset, the best performance was obtained with $w_p = 2$ and $w_{np} = 1$. From the experimental results that are shown in Table 3, it can be seen that our proposed method performs better than the baseline method on precision and recall.

	Precision	Recall
Baseline method	20.41%	19.46%
Our proposed method	30.52%	36.58%

Table 3: Performance of baseline method and our proposed method

5.3.1 Effects of the similarity threshold

As introduced in Section 5.2.3, we use a similarity threshold t to decide that two records refer to the same entity if their similarity is larger than t. Table 4 shows the performance of the baseline method and our proposed method when using five different similarity thresholds (t = 0.4, 0.5, 0.6, 0.7, 0.8), the weight of proper nouns $w_p = 2$ and the weight of non-proper nouns $w_{np} = 1$. We also illustrate these results in Figure 6.

It can be seen that the smaller t has a higher recall. On the other hand, the smaller t has a lower precision. We can also prove that our proposed method performs better than the baseline method in all similarity thresholds.

	t	0.4	0.5	0.6	0.7	0.8
Baseline method	Precision	5.49%	7.17%	20.41%	32.93%	35.29%
	Recall	45.53%	33.46%	19.46%	10.51%	9.34%
Our proposed method	Precision	6.64%	10.54%	30.52%	41.18%	42.62%
	Recall	55.64%	47.47%	36.58%	21.79%	20.23%

Table 4: Performance of baseline method and our proposed method with different similarity thresholds

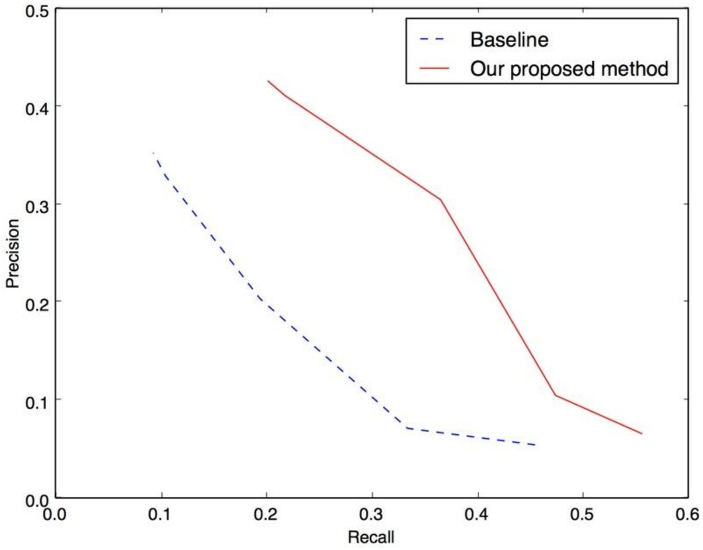


Figure 6: Precision-recall curve of baseline method and our proposed method with different similarity thresholds

5.4. Discussions

The proposed method outperforms the baseline because it correctly recognizes and transliterates the proper nouns in Japanese titles, which could help to further improve the effectiveness of cross-language record linkage. For example, the same Ukiyo-e prints with the Japanese title “目黒太鼓橋夕日の岡” and English title “Taiko Bridge, Meguro, on a Snowy Evening” are identified as true matches correctly, since the proper nouns “太鼓”, which is the name of a bridge “太鼓橋”, is recognized correctly by the proposed method, which was failed by using the baseline method.

Some proper nouns in Japanese titles are not recognized since they are not included in domain related Japanese encyclopedias, which led to fail in back-transliteration. For example, the proper noun “亀戸”, a place name, in the Japanese Ukiyo-e title “亀戸天神境内” is recognized as a non-proper noun because it is not included in Ukiyo-e related encyclopedias, which resulted to fail in back-transliteration of its corresponding transliterated word “Kameido”.

6. Conclusion

In this article, we proposed a method of proper noun recognition and transliteration, which aims at improving the accuracy of cross-language record linkage. To recognize proper nouns in Japanese metadata and obtain their transliterations, we constructed a set of Japanese and English transliterated word pairs from the English metadata, which are a part of dataset that are used in the task of cross-language record linkage. Our method employs back-transliteration to English transliterated words to acquire their original Japanese words. Experimental results have shown that this approach improved the effectiveness in cross-language record linkage between Ukiyo-e print databases in Japanese and English.

In the future, we plan to extend our method to classify the named entity type of acquired proper nouns. Besides, we plan to consider more effective similarity metrics to measure the similarity between short texts. We also plan to evaluate the effectiveness of our method on record linkage across other languages pairs and the applicability of our method to the datasets in other domains.

7. References

Batjargal, B., Kuyama, T., Kimura, F. and Maeda, A., 2014. Identifying the same records across multiple Ukiyo-e image databases using textual data in diffferent languages. In *Proceedings of the 14th ACM/IEEE Joint Conference on Digital Libraries*. pp. 193–196.

Bilac, S. and Tanaka, H., 2004. A hybrid back-transliteration system for Japanese. In *Proceedings of the 20th international conference on Computational Linguistics*, pp. 597–603.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P. and Fienberg, S, 2003. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23.

Durrani, N. Sajjad, H., Hoang, H. and Koehn, P., 2014. Integrating an unsupervised transliteration model into statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 148–153, Gothenburg, Sweden.

Elmagarmid, A., Ipeirotis, P. and Verykios, V., 2007. Duplicate record detection: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1): 1–16.

Fellegi, I.P. and Sunter, A.B., 1969. A theory for record linkage. *Journal of the American*

Statistical Association, 64(328): 1183–1210.

Huang, F. and Voge, S., 2002. Improved named entity translation and bilingual named entity extraction. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*. pp. 253–258, Pittsburgh, PA.

Jeong, K.S., Myaeng, S. H., Lee, J. S. and Choi, K. S., 1999. Automatic identification and back-transliteration of foreign words for information retrieval. *Information Processing and Management*, 35(4): 523–540.

Kang, B.J. and Choi, K.S., 2000. Automatic transliteration and back-transliteration by decision tree learning. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*. pp. 1135–1141.

Knight, K. and Graehl, J., 1998. Machine transliteration. *Computational Linguistics*, 24(4): 599–612.

Kudo, T., Yamamoto, K. and Matsumoto, Y., 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. pp. 230–237, Barcelona, Spain.

Li, Q., Li, H., Ji, H., Wang, W., Zheng, J. and Huang, F., 2012. Joint bilingual name tagging for parallel corpora. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. pp. 1727–1731, Maui, HI, USA.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. and McClosky, D., 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. pp. 55–60, Baltimore, Maryland.

Mayfield, J. Lawrie, D., McNamee, P. and Oard, D.W., 2011. Building a cross-language entity linking collection in twenty-one languages. In *Proceedings of the Cross Language Evaluate Forum*. pp. 3–13.

McNamee, P., Mayfield, J., Lawrie, D., Douglas W.O. and Doermann, D., 2011. Cross-language entity linking. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*. pp. 255–263.

Sarawagi, S. and Bhamidipaty, A., 2002. Interactive deduplication using active learning. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge discovery and data mining (KDD 2002)*. pp. 269–278, New York, New York, USA.

Song, Y. Kimura, T., Batjargal, B. and Maeda, A., 2016. Proper noun recognition in cross-language record linkage by exploiting transliterated words. In *Proceedings of the 2016 International Conference on Asian Language Processing (IALP)*. pp. 83–86, Tainan, Taiwan.