Proper Noun Recognition in Cross-Language Record Linkage by Exploiting Transliterated Words

Yuting Song
Taisuke Kimura
Biligaikhan Batjargal
Akira Maeda

Ritsumeikan University, Japan
Background

• The same data entity can exist in different languages across data sources.

<table>
<thead>
<tr>
<th>Database in Japanese</th>
<th>Database in English</th>
</tr>
</thead>
<tbody>
<tr>
<td>番号</td>
<td>92202743</td>
</tr>
<tr>
<td>作品名/資料名</td>
<td>富嶽三十六景 神奈川沖浪裏</td>
</tr>
<tr>
<td>作家/制作者</td>
<td>葛飾北斎/画</td>
</tr>
<tr>
<td>制作年</td>
<td>天保2年～4年</td>
</tr>
<tr>
<td>所蔵館</td>
<td>江戸東京博物館</td>
</tr>
<tr>
<td>番号</td>
<td>92202746</td>
</tr>
<tr>
<td>作品名/資料名</td>
<td>富嶽三十六景 深川万年橋下</td>
</tr>
<tr>
<td>作家/制作者</td>
<td>葛飾北斎/画</td>
</tr>
<tr>
<td>制作年</td>
<td>天保2年～4年</td>
</tr>
<tr>
<td>所蔵館</td>
<td>江戸東京博物館</td>
</tr>
<tr>
<td>番号</td>
<td>08200001</td>
</tr>
<tr>
<td>作品名/資料名</td>
<td>雪月花 隅田</td>
</tr>
<tr>
<td>作家/制作者</td>
<td>葛飾北斎/画</td>
</tr>
<tr>
<td>制作年</td>
<td>[天保3年]</td>
</tr>
<tr>
<td>所蔵館</td>
<td>江戸東京博物館</td>
</tr>
</tbody>
</table>

Under the Wave off Kanagawa (Kanagawa oki nami ura), also known as The Great Wave, from the series Thirty-six Views of Mount Fuji (Fugaku sanjūrokkei) Katsushika Hokusai (Japanese, Tokyo (Edo) 1760–1849 Tokyo (Edo))

Date: ca.1830–32
Medium: Polychrome woodblock print; ink and color on paper

Snow on the Sumida River (Sumida), from the series, Snow, Moon, and Flowers (Setsugekka) Katsushika Hokusai (Japanese, Tokyo (Edo) 1760–1849 Tokyo (Edo))

Date: ca.1833
Medium: Polychrome woodblock print; ink and color on paper
Motivation

- Cross-language record linkage
  - Record linkage is to find record pairs that refer to the same entity across multiple data sources
  - across multiple data sources in different languages
  - It provides opportunities for people to access multilingual information

The metadata values of records in source language

Translating metadata values

The translated metadata values in target language

Record pair comparison

Data source in target language

The metadata values of records in target language

Data source in source language

The metadata values of records in source language
Problem statement

- **Proper nouns** in metadata values are more easily to be matched
- **Correct recognition and translation** of proper nouns in metadata values could have a positive effect on cross-language record linkage
Problem statement

• **Proper nouns** in metadata values are more easily to be matched
• **Correct recognition and translation** of proper nouns in metadata values could have a positive effect on cross-language record linkage

**How to recognize proper nouns in metadata values?**

• **Named Entity Recognition system**
  • It does not perform well on metadata values
    • The metadata values are usually short texts that can’t provide sufficient grammar and syntax information
Proposed method

- Our proposed method is inspired by the following observation: **the English translation of a Japanese proper noun is usually a transliterated word**
- English transliterated words are easy to be identified
- **Back-transliteration**: converting English transliterated words to their corresponding Japanese words
Overall process

The metadata values of records in Japanese

Proper noun recognition

Translating metadata values

The translated metadata values in English

The metadata values of records in English

Identification of transliterated word

Step.1

Back-transliteration

Step.2

Post-processing

Step.3

Acquired Japanese words and their English transliterations

Step.4

Record pair comparison
Identification of transliterated word

Step.1  Identification of transliterated word

Step.2  Back-transliteration

Step.3  Post-processing

Step.4  Acquired Japanese words and their English transliterations

We assume that any word that is not listed in an English dictionary is a transliterated word

Proper noun recognition

The metadata values of records in Japanese

Proper noun recognition

Translating metadata values

The translated metadata values in English

Record pair comparison
Back-transliteration

The metadata values of records in Japanese

Step 1: Identification of transliterated word

Step 2: Back-transliteration

Step 3: Post-processing

Step 4: Acquired Japanese words and their English transliterations

To convert transliterated words to Japanese words

Proper noun recognition

Translating metadata values

The translated metadata values in English

Record pair comparison

Proper noun recognition
Back-transliteration

**English transliterated words**
- Sumida
- Fukagawa

**The phrases and their pronunciations**
- 寅田川 ↔ すみだがわ
- 亦深川 ↔ またふかがわ

**Converting transliterated words to Japanese hiragana sequences**
- Sumida → すみだ
- Fukagawa → ふかがわ

**Converting hiragana sequences to corresponding Japanese words**
- Sumida → すみだ → 隅田
- Fukagawa → ふかがわ → 深川

**Character alignment with Kanji dictionary**

**Candidate Japanese words and their transliterated words**
- 寅田 --- Sumida
- 深川 --- Fukagawa
Post-processing

The metadata values of records in Japanese

Proper noun recognition

Translating metadata values

The translated metadata values in English

The metadata values of records in English

Identification of transliterated word

Back-transliteration

Post-processing

Step.1

Step.2

Step.3

Step.4

Acquired Japanese words and their English transliterations

Record pair comparison
Post-processing

• One English transliterated words could be back-transliterated to one or more Japanese words
  • Some of them are not proper nouns
• Post-processing is to remove the words that are not proper nouns

Sano → さの →狭の、佐野
… … …

Candidate proper noun and their transliterated words
狭の ↔ Sano
佐野 ↔ Sano
… … …

Adding the constraint on the appearance of kana at the end of Japanese proper nouns
Experiments

• Linking the same Ukiyo-e records between databases in Japanese and English

• Experimental data
  • Ukiyo-e prints
    • Japanese traditional woodblock printing
    • These prints have been digitized and exhibited in many digital libraries and museums with textual metadata values in various languages

• Dataset
  • The titles of Ukiyo-e prints are used to identify the same records

<table>
<thead>
<tr>
<th>Language</th>
<th>Ukiyo-e database</th>
<th>Number of Ukiyo-e prints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese</td>
<td>Edo-Tokyo Museum</td>
<td>242</td>
</tr>
<tr>
<td>English</td>
<td>Metropolitan Museum of Art</td>
<td>3456</td>
</tr>
</tbody>
</table>

• Each Japanese title has at least one corresponding English title
• Among the 242 Japanese titles, 209 titles contain at least one proper noun (Target titles)
Cross-language record linkage

- **Recognition of proper nouns** in metadata values
  - **Our proposed method**
  - **Baseline**: MeCab – a Japanese part-of-speech and morphological analyzer
    - An example of using MeCab to recognize proper nouns

  **Input**: 深川万年橋下
  **Output**:

<table>
<thead>
<tr>
<th>Word</th>
<th>Part-of-speech</th>
<th>Sub-class of part of speech</th>
<th>Pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>深川</td>
<td>名詞-固有名詞-地名-一般</td>
<td>固有名詞</td>
<td>フカガワ</td>
</tr>
<tr>
<td>万</td>
<td>名詞-数詞</td>
<td>数詞</td>
<td>マン</td>
</tr>
<tr>
<td>年</td>
<td>名詞-普通名詞-助数詞可能</td>
<td>普通名詞</td>
<td>ネン</td>
</tr>
<tr>
<td>橋下</td>
<td>名詞-普通名詞-一般</td>
<td>普通名詞</td>
<td>キョーカ</td>
</tr>
</tbody>
</table>
Experimental results (1)

- Linking the same Ukiyo-e records between databases in Japanese and English

![Graph showing accuracy for baseline and our method across different top-n positions.](attachment:graph.png)
Experimental results (2)

- Linking the same Ukiyo-e records that have target titles
- Target titles contain at least one proper noun
Discussion

• Successful example

Proper noun is recognized correctly.

Meguro, Taiko bridge, evening, hill

Matching

Taiko Bridge, Meguro, on a Snowy Evening


• Unsuccessful example

亀戸 (Kameido) is not recognized as a proper noun.

亀戸天神境内

Translating

turtle, door, Tenjin, precinct

Matching ✗

In the Kameido Tenjin Shrine Compound

English title

Japanese title
Conclusion

• Recognizing and transliterating the proper nouns in cross-language record linkage effectively

• In the future, we plan to extend our method to classify the named entity type of acquired proper nouns
Thank you!