Crossing of Two Linguistic Ontologies: WordNet-anchored Comparison of Chinese-Japanese Kanji Words

Chu-Ren Huang*, Hotani Chiyo*, Tzu-Yi Kuo*, Shu-Kai Hsieh**
*Institute of Linguistics, Academia Sinica, Taiwan
**National Taiwan Normal University, Taiwan
Outline

- Introduction
- Previous works
- Required Resources
- Procedure and the Results
  - Hanzi (Chinese characters) Mapping
  - Finding Synonyms (Word Relations)
  - Unknown Relation Analysis
- Conclusion
Introduction

• Chinese and Japanese are two typologically different languages sharing the same orthography

• Faux amis: Challenges for meaning-based approach to Kanji/Hanzi mapping. e.g., character 湯 means ('hot soup') in Chinese, but ('hot spring') in Japanese

• Unified lexical resources are necessary in advanced multilingual knowledge processing
Introduction

• To examine and analyse the form-meaning of Hanzi/Kanji with their semantic relations through the Chinese WordNet and the Japanese Electronic Dictionary.

• By the alignment of CWN via form-meaning mappings of Japanese and Chinese words, this work may facilitate the creation of the Japanese WordNet.
Previous Works

• WordNet-like lexical Knowledge Base for Chinese include: HowNet, Chinese Concept Dictionary (CCD), and Chinese Wordnet (Huang et al. 2006).

• Character-based approach: Wong and Pala (2002); Hantology (Chou 2005); HanziNet (Hsieh 2006).

• Kanji – Hanzi mapping: no systematic linking anchored by WordNet-like lexical knowledge base.
To perform a character-based, sense-anchored comparison of Chinese and Japanese words, we employed three resources:

- **EDR** (Electronic Dictionary Research)
  - contains list of 325454 Japanese words (jwd) and their descriptions
- **CWN** (Chinese WordNet)
  - contains list of 8624 Chinese words (cwd) with glosses, synset mapping to PWN, and relations
List of Character Variants: Due to the long period of development, Chinese character system has resulted in a small set of glyph variants.

e.g., a character with the basic meaning of 'elder sister' are represented by two glyph variants as follows:

\[
\begin{align*}
\text{姊} & \quad \text{ref= "U+59CA"} \\
\text{姉} & \quad \text{ref= "U+59C9"}
\end{align*}
\]

In this study, we use a list of 125 pairs of Japanese and Chinese character variants compiled by C. Wittern (Kyoto University).
Procedure I: Hanzi Mapping

- Hanzi Mapping: Each jwd is mapped to the corresponding cwd according to their Hanzi similarity. Such mapping pairs are divided into three groups:

  1. Identical Hanzi Sequence Pairs. E.g., 頭
  2. Different Hanzi Order Pairs
     E.g. Japanese Chinese
         律法 法律
(3) Partly Identical Pairs

E.g. Japanese
相合

Chinese
相對於
合力
相形之下

After the mapping procedure, those jwd and cwd that are not mapped are classified into (4) independent Japanese group, and (5) independent Chinese group, respectively.
Hanzi Mapping: the Results

Table 1) J-C Hanzi Similarity Distribution

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of Words</th>
<th>Number of J-C Word Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Identical Hanzi Sequence Pairs</td>
<td>2815 jwds</td>
<td>20199</td>
</tr>
<tr>
<td>(2) Different Hanzi Order Pairs</td>
<td>204 jwds</td>
<td>473</td>
</tr>
<tr>
<td>(3) Partly Identical Pairs</td>
<td>264917 jwds</td>
<td>8438099</td>
</tr>
<tr>
<td>(4) Independent Japanese</td>
<td>57518 jwds</td>
<td>-</td>
</tr>
<tr>
<td>(5) Independent Chinese</td>
<td>851 cwds</td>
<td>-</td>
</tr>
</tbody>
</table>
Procedure II: Finding Synonyms

• Finding Synonymes: after the character-based mapping, pairs in (1) and (2) are divided into three subgroups.

(1-1, 2-1) Synonym pairs with identical POS

E.g. (1-1) 以降 : afterwards (noun)
(2-1) 兄弟 (Japanese) and 弟兄 (Chinese): brother (noun)
Procedure II: Finding Synonyms

(1-2, 2-2) Synonym pairs with unmatched POS
E.g. (1-2) 意味 : sense (noun in EDR and verb in CWN)
(2-2) 定規 (Japanese) and 規定 (Chinese) :
rule (noun in EDR and no POS is indicated in CWN)

(1-3, 2-3) Unknown relation
E.g. Japanese Chinese
(1-3) 灰 : ash (noun) 灰 : dust (no POS indicated)
(2-3) 愛心 : affection (noun) 心愛 : dear, darling
(no POS indicated)
Procedure II: Finding Synonyms

- To find the relation of J-C word pairs

1. Check if translations have any match
   - Yes: Synonym with the same POS pairs
   - No: Unknown relation

2. Check if the POS are identical
   - Yes: Synonym with the same POS pairs
   - No: Synonym with unmatched POS pairs
# Finding Synonyms: the Results

Table 2) Identical Hanzi Sequence Pairs (20199 pairs) Synonymous Relation Distribution

<table>
<thead>
<tr>
<th></th>
<th>Number of 1-to-1 Form-Meaning Pairs Found by Machine Processing (% in (1))</th>
<th>Number of 1-to-1 Form-Meaning Pairs Found by Manual Analysis (% in (1))</th>
<th>* Number of Many-to-Many Form-Meaning Pairs Found by Manual Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1-1) Synonym with the same POS pairs</td>
<td>92 (0.5%)</td>
<td>35 (0.2%)</td>
<td>26</td>
</tr>
<tr>
<td>(1-2) Synonym with unmatched POS pairs</td>
<td>425 (2.1%)</td>
<td>252 (1.2%)</td>
<td>150</td>
</tr>
<tr>
<td>(1-3) unknown relation</td>
<td>19682 (97.4%)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
## Procedure and the Result

### Table 3) Identical Hanzi But Different Order Pairs (505 pairs) Synonymous Relation Distribution

<table>
<thead>
<tr>
<th><strong>(2-1) Synonym with the same POS pairs</strong></th>
<th>Number of 1-to-1 Form-Meaning Pairs Found by Machine Processing (% in (2))</th>
<th>Number of 1-to-1 Form-Meaning Pairs Found by Manual Analysis (% in (2))</th>
<th>* Number of Many-to-Many Form-Meaning Pairs Found by Manual Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(2-2) Synonym with unmatched POS pairs</strong></td>
<td>14 (3.0%)</td>
<td>11 (2.3%)</td>
<td>10</td>
</tr>
<tr>
<td><strong>(2-3) unknown relation</strong></td>
<td>459 (97.0%)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Procedure III: Unknown Relation Analysis

- Unknown Relation Analysis
  - the pairs with unknown relation are divided into the following four different groups

  (A) Only comparison info. of jwd is missing
  E.g.
  (1-3-A) No English translation for 足 in EDR
  (2-3-A) No English translation for 運命 in EDR
Procedure III: Unknown Relation Analysis

(B) Only comparison info. of cwd is missing
E.g. (1-3-B) No English translation nor synset for 有無 in CWN
(2-3-B) No English translation nor synset for 明星 in CWN

(C) No comparison info. is missing
E.g. Japanese Chinese
(1-3-C) 火力 : firepower (noun) power, powerfulness (no POS)
Procedure III: Unknown Relation Analysis

(D) Both comparison info. are missing
E.g. (1-3-D) No English translation nor synset for 機動 in both EDR and CWN
(2-3-D) No English translation nor synset for 山中 in EDR and for 中山 in CWN
Procedure III: Unknown Relation Analysis

- Sort group (A), (B) and (C) into possible synonym pairs and non-synonym pairs

A) Check if the definition of jwd contains any of the ctranslations or cwd synset

B) Check if the definition of cwd contains jtranslation

C) Do both the methods that for (A) and (B)
### Procedure III and the Result

Table 4) Identical Hanzi Sequence Pairs with Unknown Relation (19682 pairs) distribution

<table>
<thead>
<tr>
<th></th>
<th>Number of Pairs (% in 1-3)</th>
<th>Number of Possible Synonym Pairs (% in 1-3)</th>
<th>Number of Non-Synonym Pairs (% in 1-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Missing the Japanese translation</td>
<td>8428 (42.8%)</td>
<td>590 (3.0%)</td>
<td>7838 (39.8%)</td>
</tr>
<tr>
<td>(B) Missing Chinese translation and the synset</td>
<td>2275 (11.6%)</td>
<td>0 (0.0%)</td>
<td>2275 (11.6%)</td>
</tr>
<tr>
<td>(C) No missing information</td>
<td>5720 (29.1%)</td>
<td>296 (1.5%)</td>
<td>5424 (27.6%)</td>
</tr>
<tr>
<td>(D) Missing both translations and the synset</td>
<td>3259 (16.6%)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 5: Identical Hanzi But Different Order Pairs with Unknown Relation (485 pairs)

distribution

<table>
<thead>
<tr>
<th>Description</th>
<th>Number of Pairs (% in 2-3)</th>
<th>Number of Possible Synonym Pairs (% in 2-3)</th>
<th>Number of Non-Synonym Pairs (% in 2-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Missing the Japanese translation</td>
<td>199 (43.7%)</td>
<td>5 (1.1%)</td>
<td>194 (42.6%)</td>
</tr>
<tr>
<td>*** (B) Missing Chinese translation and the synset</td>
<td>46 (9.5%)</td>
<td>0 (0.0%)</td>
<td>46 (9.5%)</td>
</tr>
<tr>
<td>(C) No missing information</td>
<td>151 (32.9%)</td>
<td>10 (2.2%)</td>
<td>141 (30.7%)</td>
</tr>
<tr>
<td>(D) Missing both translations and the synset</td>
<td>63 (13.0%)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Conclusion

- More than 75% of pairs are found to be non-synonyms, the majority of pairs are faux amis.

<table>
<thead>
<tr>
<th>Machine Analysis</th>
<th>Pairs Found to be Synonym</th>
<th>Pairs Found to be Non-Synonym</th>
<th>Unknown Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>% in (1)</td>
<td>6.9%</td>
<td>76.9%</td>
<td>16.1%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Including Manual Analysis</th>
<th>Pairs Found to be Synonym</th>
<th>Pairs Found to be Non-Synonym</th>
<th>Unknown Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>% in (1)</td>
<td>5.8%</td>
<td>78.1%</td>
<td>16.1%</td>
</tr>
</tbody>
</table>
Conclusion

Table 7) Identical Hanzi But Different Order Pairs (473 pairs) Lexical Semantic Relation

<table>
<thead>
<tr>
<th></th>
<th>Pairs Found to be Synonym</th>
<th>Pairs Found to be Non-Synonym</th>
<th>Unknown Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Analysis</td>
<td>29</td>
<td>381</td>
<td>63</td>
</tr>
<tr>
<td>% in (2)</td>
<td>6.1%</td>
<td>80.5%</td>
<td>13.3%</td>
</tr>
<tr>
<td>Including Manual Analysis</td>
<td>26</td>
<td>384</td>
<td>63</td>
</tr>
<tr>
<td>% in (2)</td>
<td>5.5%</td>
<td>81.2%</td>
<td>13.3%</td>
</tr>
</tbody>
</table>
Conclusion

• However, it is not certain whether the pairs are really non-synonyms and what their actual semantic relations are.

• In the further experiment, we will try to find the semantic relations of those pairs found to be non-synonyms pairs.
THANK YOU