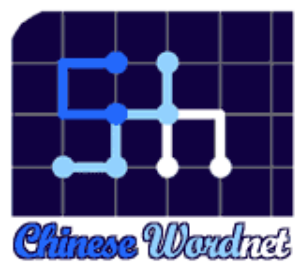


Crossing of Two Linguistic Ontologies: WordNet-anchored Comparison of Chinese-Japanese Kanji Words

Chu-Ren Huang*, Hotani Chiyo*, Tzu-Yi Kuo*, Shu-Kai Hsieh**

*Institute of Linguistics, Academia Sinica, Taiwan

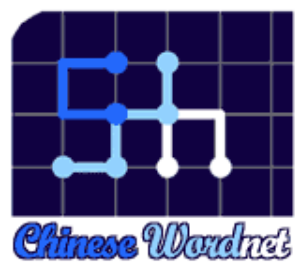
**National Taiwan Normal University, Taiwan



Outline

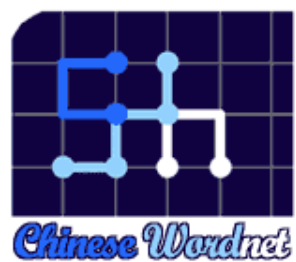


- Introduction
- Previous works
- Required Resources
- Procedure and the Results
 - Hanzi (Chinese characters) Mapping
 - Finding Synonyms (Word Relations)
 - Unknown Relation Analysis
- Conclusion



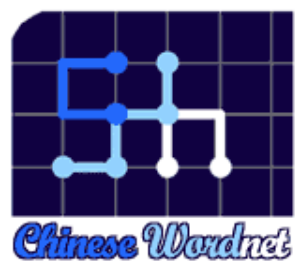
Introduction

- 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



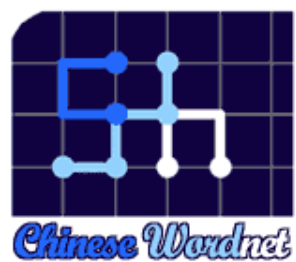
Introduction

- 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

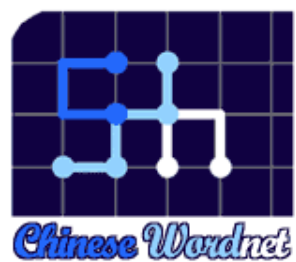
- WordNet-like lexical Knowledge Base for Chinese include: **HowNet**, **Chinese Concept Dictionary (CCD)**, and **Chinese Wordnet** (Huang et al. 2006).
- Character-based approach: Wong and Pala (2002); **Hantology** (Chou 2005); **HanziNet** (Hsieh 2006).
- Kanji – Hanzi mapping: no systematic linking anchored by WordNet-like lexical knowledge base.



Required Resources: EDR/CWN

To perform a character-based, sense-anchored comparison of Chinese and Japanese words, we employed three resources:

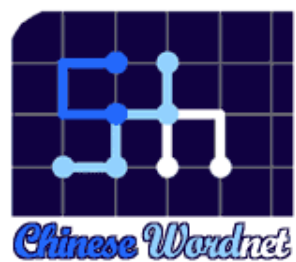
- **EDR** (Electronic Dictionary Research)
 - contains list of 325454 Japanese words (jwd) and their descriptions
- **CWN** (Chinese WordNet)
 - contains list of 8624 Chinese words (cwd) with glosses, synset mapping to PWN, and relations



Required Resources: List of Character Variants

- **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:

```
<c type="tw" ref="U+59CA"> 姊 </c>  
<c type="jp" ref="U+59C9"> 姉 </c>
```
- 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

Procedure I: Hanzi Mapping

(3) Partly Identical Pairs

E.g. Japanese

相合

Chinese

相對於

合力

相形之下

語音

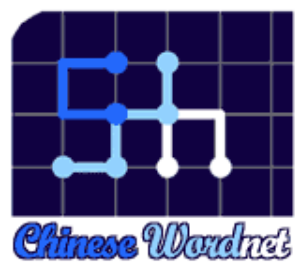
After the mapping procedure, those jwd and cwd that are not mapped are classified into (4) independent Japanese group, and (5) independent Chinese group, respectively.

Hanzi Mapping: the Results



Table 1) J-C Hanzi Similarity Distribution

	Number of Words	Number of J-C Word Pairs
(1) Identical Hanzi Sequence Pairs	2815 jwds	20199
(2) Different Hanzi Order Pairs	204 jwds	473
(3) Partly Identical Pairs	264917 jwds	8438099
(4) Independent Japanese	57518 jwds	-
(5) Independent Chinese	851 cwds	-



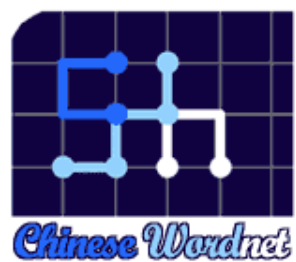
Procedure II: Finding Synonyms

- 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

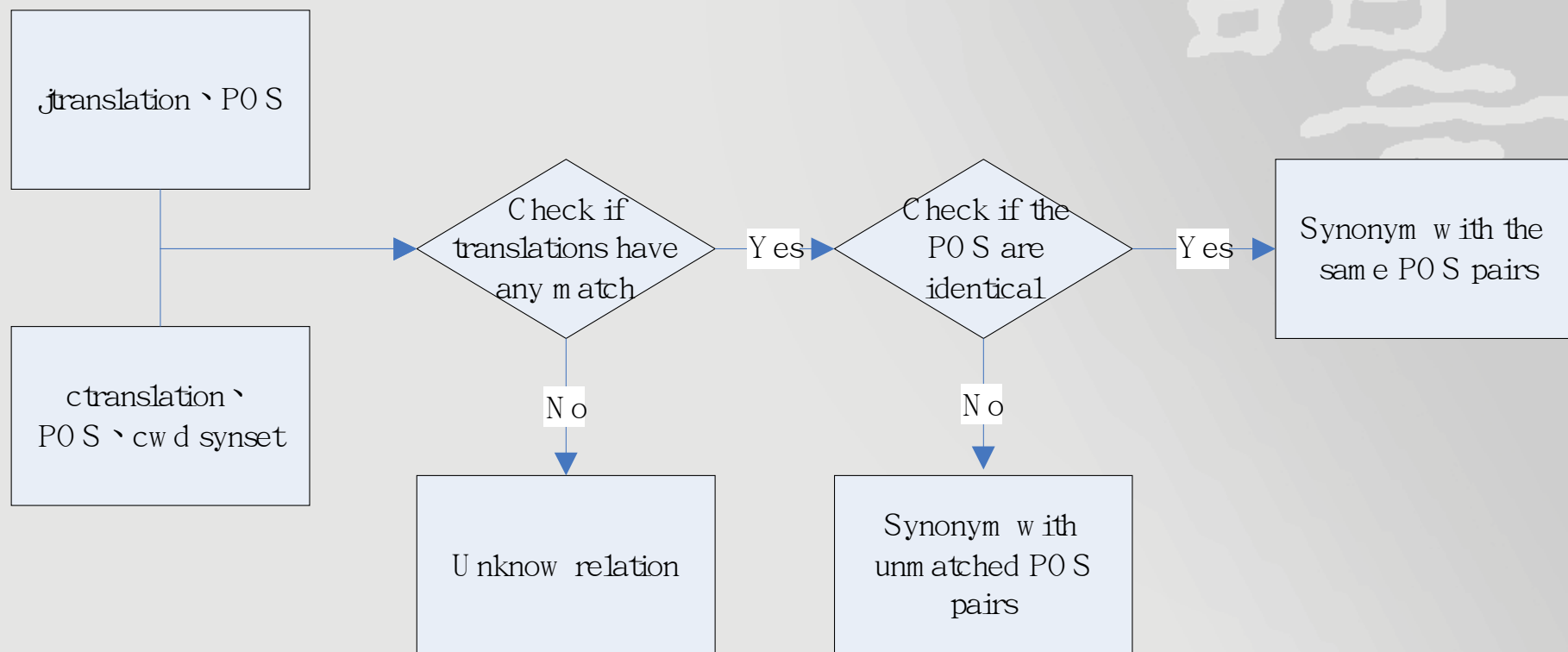
E.g. Japanese Chinese

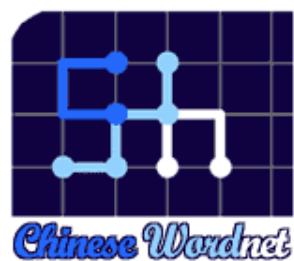
(1-3) 灰 : ash (noun) 灰 : dust (no POS indicated)

(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





Finding Synonyms: the Results

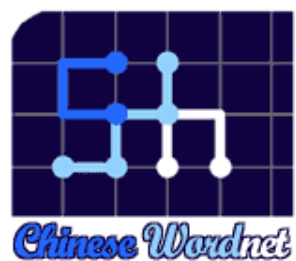
Table 2) Identical Hanzi Sequence Pairs (20199 pairs) Synonymous Relation Distribution

	Number of 1-to-1 Form-Meaning Pairs Found by Machine Processing (% in (1))	Number of 1-to-1 Form-Meaning Pairs Found by Manual Analysis (% in (1))	* Number of Many-to-Many Form-Meaning Pairs Found by Manual Analysis
(1-1) Synonym with the same POS pairs	92 (0.5%)	35 (0.2%)	26
(1-2) Synonym with unmatched POS pairs	425 (2.1%)	252 (1.2%)	150
(1-3) unknown relation	19682 (97.4%)	-	-

Procedure and the Result

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

	Number of 1-to-1 Form-Meaning Pairs Found by Machine Processing (% in (2))	Number of 1-to-1 Form-Meaning Pairs Found by Manual Analysis (% in (2))	* Number of Many-to-Many Form-Meaning Pairs Found by Manual Analysis
** (2-1) Synonym with the same POS pairs	0 (0.0%)	0 (0.0%)	0
(2-2) Synonym with unmatched POS pairs	14 (3.0%)	11 (2.3%)	10
(2-3) unknown relation	459 (97.0%)	-	-



Procedure III: Unknown Relation Analysis

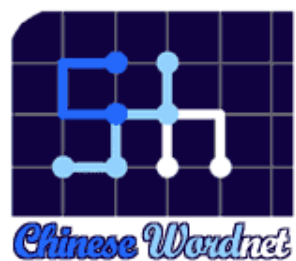
- Unknown Relation Analysis
 - the pairs with unknown relation are divided into the following four different groups

(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

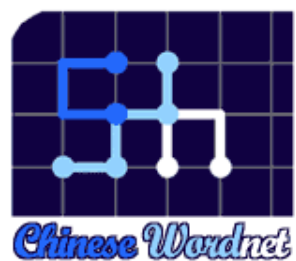


Procedure III: Unknown Relation Analysis

(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

	Number of Pairs (% in 1-3)	Number of Possible Synonym Pairs (% in 1-3)	Number of Non-Synonym Pairs (% in 1-3)
(A) Missing the Japanese translation	8428 (42.8%)	590 (3.0%)	7838 (39.8%)
*** (B) Missing Chinese translation and the synset	2275 (11.6%)	0 (0.0%)	2275 (11.6%)
(C) No missing information	5720 (29.1%)	296 (1.5%)	5424 (27.6%)
(D) Missing both translations and the synset	3259 (16.6%)	-	-

Procedure III and the Result

Table 5) Identical Hanzi But Different Order Pairs with Unknown Relation (485 pairs) distribution

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

Conclusion

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

Table 6) Identical Hanzi Sequence Pairs (20199 pairs) Lexical Semantic Relation

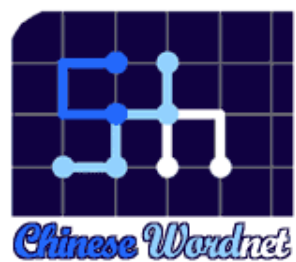
	Pairs Found to be Synonym	Pairs Found to be Non-Synonym	Unknown Relation
Machine Analysis	1403	15537	3259
% in (1)	6.9%	76.9%	16.1%
Including Manual Analysis	1173	15767	3259
% in (1)	5.8%	78.1%	16.1%

Conclusion



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

	Pairs Found to be Synonym	Pairs Found to be Non-Synonym	Unknown Relation
Machine Analysis	29	381	63
% in (2)	6.1%	80.5%	13.3%
Including Manual Analysis	26	384	63
% in (2)	5.5%	81.2%	13.3%



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.

語音

THANK YOU