WordNet Based Comparison of Language Variation: A Study Based on CCD and CWN

Jia-Fei Hong
Institute of Linguistics, Academia Sinica
Institute of Linguistics, Academia Sinica, Nankang, Taipei, Taiwan 115
jiafei@gate.sinica.edu.tw

Chu-Ren Huang
Institute of Linguistics, Academia Sinica
Institute of Linguistics, Academia Sinica, Nankang, Taipei, Taiwan 115
churen@sinica.edu.tw

Yang Liu
Institute of Computational Linguistics, Peking University
Peking, China
liuyang@pku.edu.cn

Abstract

This paper will deal with the lexica of comparing the Chinese Concept Dictionary (CCD) with the Chinese WordNet (CWN) by WordNet. CCD is a WordNet-like semantic lexicon that developed by the Institute of Computational Linguistics, Peking University. And CWN is a bilingual wordnet by linking to the SUMO ontology that developed by Academia Sinica Bilingual Ontological WordNet. In this paper, we will base on WordNet database to show several situations for both CCD and CWN, such as: the same translation for them, zero translation only for CCD or CWN, and unique translation only for CCD or CWN. Then, through these analyses, we could find out the unique usage of English translating for traditional Chinese Characters or simplified Chinese characters.

Keywords: CCD; CWN; Sense; Concept; Synset; WordNet; Bilingual

Introduction

Miller thought that they could use the synonym sets to represent the lexicon concepts and describe the lexicon contents, so they established the WordNet. Recently, there are many research teams to deal with translations by the knowledge base of WordNet. In the bilingual Chinese-English WordNet, there are two teams to set about analyzing these data.

Regard translating as the foundation, the establishment of the reliability and relative problems for a bilingual wordnet. In these data, we could put emphasis on comparing with the coincidence and their analyses, and two different translated contrasts. Finally, how greatly we could find the one new correction to help that is translated to the mistake of correcting.

Regard English WordNet as the database (English is as intermediary's language), we could set up and implement the concept of the corresponding knowledge base of lexica for traditional Chinese Characters or simplified Chinese characters.

1 WordNet and Sinica BOW

WordNet, an electronic lexical database, is considered to be one of the most important resources available to researchers in computational linguistics, text analysis, and many related areas (Miller et al., 1993; Fellbaum, 1998). Its design is inspired by current psycholinguistic and computational theories of human lexical memory. English nouns, verbs, adjectives, and adverbs are organized into synonym sets, each representing one underlying lexicalized concept. Different semantic relations link the synonym sets (synsets).

There are several versions of WordNet, with WordNet 2.0 being the most recent one. The differences between these versions include the quantity of synsets and their definition. The version of WordNet that we use in this research is version 1.6, since this is the version most widely used by computational linguists. There are nearly 100,000 synsets in this version.

We mentioned earlier that we adopted the bilingual domain taxonomy to increase the versatility of our domain processing. Similarly, we use a bilingual wordnet as our lexical knowledgebase to achieve bilingual support to our study at the

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 61–68.

lexico-conceptual level. Each English synset was given up to 3 most appropriate Chinese translation equivalents. And in cases where the translation pairs are not synonyms, their semantic relations are marked (Huang et al. 2003). The resulted bilingual wordnet is further linked to the SUMO ontology to form the Academia Sinica Bilingual Ontological Wordnet (Sinica BOW, Huang and Chang, 2004). We use the semantic relations in bilingual resource to expand and predict domain classification when it cannot be judged directly from a lexical lemma.

2 WordNet and CCD

WordNet, an electronic lexical database, is considered to be one of the most important resources available to researchers in computational linguistics, text analysis, and many related areas (Miller et al., 1993; Fellbaum, 1998). Its design is inspired by current psycholinguistic and computational theories of human lexical memory. English nouns, verbs, adjectives, and adverbs are organized into synonym sets, each representing one underlying lexicalized concept. Different semantic relations link the synonym sets (synsets).

There are several versions of WordNet, with WordNet 2.0 being the most recent one. The differences between these versions include the quantity of synsets and their definition. The version of WordNet that we use in this research is version 1.6, since this is the version most widely used by computational linguists. There are nearly 100,000 synsets in this version.

We mentioned earlier that we adopted the bilingual domain taxonomy to increase the versatility of our domain processing. Similarly, we use a bilingual wordnet as our lexical knowledgebase to achieve bilingual support to our study at the lexico-conceptual level. Each English synset was given up to 3 most appropriate Chinese translation equivalents. And in cases where the translation pairs are not synonyms, their semantic relations are marked (Huang et al. 2003). The resulted bilingual wordnet is further linked to the SUMO ontology to form the Academia Sinica Bilingual Ontological Wordnet (Sinica BOW, Huang and Chang, 2004). We use the semantic relations in bilingual resource to expand and predict domain classification when it cannot be judged directly from a lexical lemma.

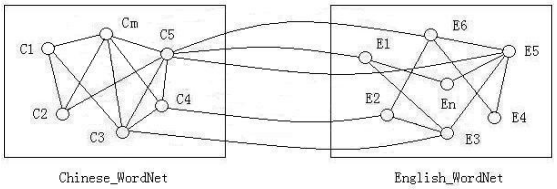
3 WordNet and CCD

CCD is a bilingual Chinese_English WordNet from the frame of WordNet. Compatible with WordNet on the CCD specification, the research team describes these senses under the prerequisite that namely does not destroy WordNet frame with the synonym defining concepts and relationships. On the other hand, they also consider that there may exist different descriptive structures between Chinese and English, so CCD put to stress the characteristic of Chinese to not only express the language contents of Chinese but also certainly develop the relationships between the content and the concept for Chinese.

The research team focuses on the structure of CCD, which presents a concept defined by a synonyms set (Synset) and a network of concepts based on the hypernymy hierarchy, the basic relationship, with other supplementary relationships. The deductive rules on this semantic network are mathematically formalized, which could be well applied to Chinese Semantic analysis.

From 2000/09, in Peking University, the Institute of Computational Linguistics already were based on WordNet to research CCD for establishing the bilingual Chinese_English WordNet, which supply several services such as the direct machine translation(MT), information extracting(IE)... and so on.

Owing to the concern of two different knowledge base and their concepts, there are pretty complex in the inner structure of CCD. CCD includes the great scope and complex structure of the pair sub-network that there are about 10^5 concept nodes and 10^6 concept relationships for each sub-network This relationship will shown as below:



Firuge1. The complex relationship structure of sub-network

4 Analysis

4.1 Translation analysis

4.1.1 The zero translation for CCD and CWN

Base on WordNet database, there is no zero translation for CCD and CWN in the same synset. The statistics will shown as below:

Title	Frequency
The lackness translation for CCD and CWN	**0**

Table 1 The zero translation for CCD and CWN

4.1.2 The zero translation only for CCD

In CCD database, there are many synsets that were not dealt with their translations, because maybe the Chinese translation refers to unknown object or maybe the Chinese translation present unknown word. Thus, there are about 4000 synsets of zero translation for CCD. The detail information will be exhibited as table2 and table3:

Title	Frequency
The lackness translation only for CCD	**4000**

Table 2 The zero translation only for CCD

Mean	Synset	The same translation for CCD and CWN	The same translation for CCD and CWN	The same translation for CCD and CWN
the craft of making fireworks	pyrotechnics,pyrotechny	X	X	煙火製造術 (yan huo zhi zao shu)
attack with teargas; subject to teargas fumes	teargas	X	X	向…投催淚彈 (xiang … tou cui lei dan), 施放催淚瓦斯 (shi fang cui lei wa si)

Table 3 The examples: The zero translation only for CCD

4.1.3 The zero translation only for CWN

Based on WordNet database, we translate all synsets that we must assign the applicable terms for them, regardless of words, unknown words, phrases, utterances, sentences and so on. Anyway, we must do our best to translate all synsets, so in CWN, there is no zero translation.

Title	Frequency
The lackness translation only for CWN	**0**

Table 4 The zero translation only for CWN

4.1.4 The same translation for CCD and CWN

Well, CCD and CWN all depend on WordNet to establish the bilingual Chinese_English WordNet, surely, they must have many similar cases that the details are shown such as below:

Title	The number of the same translation for CCD and CWN							
Category	1	2	3	4	5	6	7	**Total**
Frequency	10007	3674	1133	63	17	4	1	**14899**

Table 5 The same translation for CCD and CWN

Title	The POS of the same translation for CCD and CWN	
Category	**The number of word**	**Frequency**
N	1	9438
N	2	3140
N	3	907
…	…	…
V	1	419
V	2	446
V	3	187
…	…	…
R	1	55
R	2	51
R	3	26
…	…	…
A	1	95
A	2	37
A	3	13
…	…	…
Total		**14899**

Table 6 Statistics: The POS of the same translation for CCD and CWN

Mean	**Synset**	**The same trans-lation for CCD and CWN**	**The unique trans-lation for CCD**	**The unique trans-lation for CWN**
a shelf on which to keep books	bookshelf	書架 (shu jia) 書櫃 (shu gui) 書櫥 (shu chu)	X	X
text that is typed or printed on paper	hard_copy	硬複本 (ying fu ben) 硬拷貝 (ying kao bei) 硬式複本 ((ying shi fu ben) 硬性複本 ((ying xing fu ben)	X	X

Table 7 The examples: The same translation for CCD and CWN

4.1.5 The unique translation for CCD

In spite of many similar translations for CCD and CWN, several translations are still unique only for CCD. The statements will be presented as below:

The number of the unique translation for CCD	
Category	**Frequency**
1	41018
2	19569
3	10164
4	4650
5	1698
6	854
7	432
8	298
9	197
10	152
11	103
12	63
13	50

14	41
15	21
17	20
16	14
19	14
18	13
21	8
22	8
26	6
20	6
23	2
25	2
33	1
29	1
28	1
34	1
45	1
Total	**79408**

Table 8 The unique translation for CCD

Title	**The POS of the same translation for CCD**	
Category	**The number of word**	**Frequency**
N	1	31890
N	2	9444
N	3	3333
…	…	…
V	1	4500
V	2	2313
V	3	1338
…	…	…
R	1	850
R	2	1356
R	3	798
…	…	…
A	1	3778
A	2	6456
A	3	4695
…	…	…
Total		**79408**

Table 9 Statistics: The POS of the same translation for CCD

Mean	**Synset**	**The same trans-lation for CCD and CWN**	**The unique trans-lation for CCD**	**The unique trans-lation for CWN**
keep from exhaling or expelling	hold	X	屏氣 (bing qi)	摒住（呼吸）(bing zhu(hu xi))
a period of time assigned for work	hours	工作時數 (gong zuo shi shu)	課時(ke shi)	X

Table 10 The examples: The unique translation for CCD

4.1.6 The unique translation for CWN

In the same state, several translations are unique only for CWN. The statements also will be presented as below:

Title									
The number of the unique trans-lation for CWN	1	2	3	4	5	6	7	8	**Total**
Freq.	33735	9914	2806	180	48	12	9	3	**46708**

Table 11 The unique translation for CWN

Title	The POS of the same translation for CCD	
Category	**The number of word**	**Frequency**
N	1	21197
N	2	2024
N	3	241
…	…	…
V	1	2528
V	2	473
V	3	63
…	…	…
R	1	1690
R	2	1021
R	3	186
…	…	…
A	1	8320
A	2	6396
A	3	2316
…	…	…
Total		**46708**

Table 12 Statistics: The unique translation for CWN

Mean	**Synset**	**The same trans-lation for CCD and CWN**	**The unique trans-lation for CCD**	**The unique trans-lation for CWN**
manage not to throw up	keep_ down	X	咽 (yan) 咽下 (yan xia) 不吐出 (bu tu chu)	忍著不吐出 (ren zhe bu tu chu)
coat a metal with an oxide coat	anodize	電鍍 (dian du) 陽極化 (yang ji hua)	X	對…作陽極化處理 (dui…zuo yang ji hua chu li)

Table 13 The examples: The unique translation for CWN

4.1.7 The duplicate translation for CCD and CWN

Because in counting unique translation for CCD or CWN, we duplicate some synsets, there are about 41373. For instance, the synset ”aborning”, we count once time for CCD unique translation and once time for CWM, for this reason, we should take out the duplicate translation for CCD and CWN.

Title	**Frequency**
The repeated translation for CCD and CWN	**41373**

Table 14 The duplicate translation for CCD and CWN

Mean	**Synset**	**The same trans-lation for CCD and CWN**	**The unique trans-lation for CCD**	**The unique trans-lation for CWN**
in the process of being born or beginning	aborning	X	過程 (guo cheng)	在生產中 (zai sheng chan zhong)
on and on for a long time	no_end, without_stopping	X	不停 (bu ting) 大量 (bu liang)	永不停止的 (yong bu ting zhi de)

Table 15 The examples: The duplicate translation for CCD and CWN

4.2 Translative words analysis

In establishing the bilingual Chinese_English WordNet for CCD and CWN, we find out another state that translative words for them. In here, by the way, we record this appearance. The total numbers are 420515 tokens of translative words for CCD and CWN such as Table19; other records also be shown as Table16 to Table 18.

4.2.1 The number of unique translative words for CCD

Title	Token	Percentage
The number of unique translative words for CCD	**177174**	**42.13%**

Table 16 The number of unique translative words for CCD

4.2.2 The number of unique translative words for CWN

Title	Token	Percentage
The number of unique translative words for CWN	**116043**	**27.60%**

Table 17 The number of unique translative words for CWN

4.2.3 The number of the same translative words for CCD and CWN

Title	Token	Percentage
The number of the same translative words for CCD and CWN	**127298**	**30.27%**

Table 18 The number of the same translative words for CCD and CWN

4.2.4 The number of total translative words for CCD and CWN

Title	Token	Percentage
The number of total translative words for CCD and CWN	**420515**	**100.00%**

Table 19 The number of total translative words for CCD and CWN

Conclusion

As the result of several kinds of above different analyses, the two translation data are based on WordNet to deal with lexica for CCD and CWN, but the results are truly distinct. So, in this paper, our important viewpoint is that finding out the unique usage of English translating for traditional Chinese Characters or simplified Chinese characters.

References

1. Fellbaum C.. WordNet: An Electronic Lexical Database. Cambridge: MIT Press (1998).
2. Huang, Chu-Ren. Elanna I. J. Tseng, Dylan B. S. Tsai, Brian Murphy. 2003. Cross-lingual Portability of Semantic relations: Bootstrapping Chinese WordNet with English WordNet Relations. Languages and Linguistics. 4.3. (2003)509-532.
3. Huang, Chu-Ren, and Ru-Yng Chang. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO". Presented at the 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon. Portugal. 26-28 May (2004).
4. Huang Chu-Ren, Xiang-Bing Li, Jia-Fei Hong, 2004, "Domain Lexico-Taxonomy:An Approach Towards Multi-domain Language Processing", Asian Symposium on Natural Language Processing to Overcome Language Barriers, The First International Joint Conference on Natural Language Processing (IJCNLP-04) (2004).

5. Miller G. A., R. Beckwith, C. Fellbaum, D. Gross and K. Miller. 1993. “Introduction to WordNet: An On-line Lexical Database,” In Proceedings of the fifteenth International Joint Conference on Artificial Intelligence.
6. 于江生, 劉揚, 俞士汶。2003。中文概念詞典規格說明。Journal of Chinese language and Computing, 13(2) 177-194。
7. 于江生, 俞士汶。2004。中文概念詞典的結構。中文信息學報(Journal of Chinese Information Processing), vol. 16 No. 4 (2004)12-21。
8. 劉揚, 俞士汶, 于江生。2003。CCD 語義知識庫的構造研究。2003 中國計算機大會(CNCC’2003)。