LANGUAGE AND LINGUISTICS 4.3:509-532, 2003
2003-0-004-003-000048-1

Cross-lingual Portability of Semantic Relations: Bootstrapping Chinese WordNet with English WordNet Relations*

Chu-Ren Huang Elanna I. J. Tseng Dylan B. S. Tsai Brian Murphy
Academia Sinica

Establishing correspondences between wordnets of different languages is essential to multilingual knowledge processing. We claim that such correspondences must be based on lexical semantic relations, rather than top ontology or word translations. In particular, we define a translation equivalence relation as a bilingual lexical semantic relation. Such relations can then be part of the logical entailment predicting whether source language semantic relations will hold in target language or not. Our claim is tested with a study on 210 Chinese lexical lemmas and their possible semantic-relation links bootstrapped from the Princeton WordNet. The results show that lexical semantic-relation translations are indeed highly precise when they are logically inferable. The study has positive implications for bootstrapping of language wordnets with insufficient monolingual lexical resources.

Key words: (lexical) Semantic Relation, WordNet, Bilingual Semantic Relation, Translation Equivalence, Chinese WordNet

1. Introduction

WordNet databases have become essential for both lexical semantic studies and computational linguistic applications. The existing wordnets, Princeton WordNet (WN) and EuroWordNet (EWN, Vossen 1998), provide rich information for NLP applications. WN is the first formal architecture for representing a complete linguistic ontology with semantic relations (Fellbaum 1998). The fundamental relation in WN is synonymy. In

* This study was mainly supported by the NSC-NSF joint International Digital Library Project (IDLP) on Transferring Bilingual Semantic Relations. Additional support was received from a joint collaborative project between Global View Co. and the Institute of Linguistics of Academia Sinica, as well as from the Linguistic Anchoring project of Taiwan's NDAP. We would like to thank several colleagues who were involved in both infrastructure building and research idea brainstorming: Echa Chang, Keh-jiann Chen, Sue-jin Ker, and Chang-hua Yang. Kathleen Ahrens carefully read and commented on this paper. Last, but not least, the comments and suggestions of all CLSW3 participants are greatly appreciated. Any remaining errors are, of course, our own.

WN, a synset is defined as a set of synonyms. Synsets are basic nodes in the semantic network. Other semantic relations link synsets to form a hierarchical framework. The strategy that WN adopted was to propose primitive semantic classes (known as unique beginners) in each part-of-speech (POS) that can cover all lemmas within that POS category. Then they assigned semantic relations between synsets. Since they started with the top categories, it is regarded as a top-down approach.

One the other hand, the defining characteristics of EWN is its multi-lingual nature. It is an integrated system that contains eight European language sub-wordnets. Each sub-wordnet is an independent monolingual wordnet, which refers to and is linked to the English WN synset(s) via the Inter-Lingual-Index (ILI). Consequently, sub-wordnets can be linked via the ILI. What is common in these two wordnet systems is they all established top beginner classes and built the hierarchy based on those primitives. The difference between them is that EWN introduced a multilingual aspect.

Given that building wordnets is intensive work requiring rigorous semantic distinctions, neither the WN nor the EWN construction methodology seems to be efficient for future work on additional languages. First, given available wordnet resources, employing only human labor like the original WN would be wasteful and cannot guarantee compatibility. On the other hand, the simultaneous multilingual work with common resources as well as task sharing pioneered by the EWN is not practical in the absence of a shared multilingual community with a single sufficient and dependable funding source such as the European Union.

One of the most attractive alternatives for wordnets in other languages is to bootstrap from an existing wordnet based on bilingual correspondences (e.g. Pianta, et al., 2002). Such work has both theoretical and computational implications. Theoretically, it allows us to test if semantic relations can be transported cross-linguistically, and, if so, which ones. Computationally, it is an experiment in the semi-automatic construction of a monolingual wordnet via a bilingual wordnet.

In this paper, we shall discuss our construction of a sample Chinese WordNet (preCWN). This preCWN contains the 210 most frequent Chinese lemmas, according to their distribution in the Sinica Corpus. Unlike the top-ontology approach in other wordnets, we use a bottom-up strategy. That is, we first link Chinese lemmas to English WN synsets. Then, we adopt the semantic relations that have already been assigned in WN. After the automatic linking has been done, human labor is used to inspect the correctness of the bootstrapping outcome. This preCWN project has two purposes: 1) to study the retention of semantic relations across languages; 2) to build a bilingual (Chinese/English) lexical database with semantic relations specified for all translation correspondences (i.e. synonymy or otherwise).

This paper is structured as follows: after the introductory section, our methodology

and procedures will be outlined step by step in section 2. In section 3, we shall show our results and a preliminary statistical evaluation. We shall also discuss the significance of the results. In the conclusion, we shall evaluate the methodology, especially the feasibility of cross-linguistic inference of semantic relations.

2. From TEDB to preCWN

To establish a preCWN, we assign Chinese lemmas to English WN synsets based on our WN-based database of translation equivalences. The lemmas covered belong to four different POS categories: noun, verb, adjective, and adverb. The translation equivalences database (TEDB—see section 2.2) provides between one and three Chinese translation equivalents for each English synset. Since each Chinese lemma is linked to at least one corresponding synset, we can adopt the English semantic relations in WN and posit them as potential semantic relations in Chinese.[1] Then using the TEDB we obtained a new list of Chinese lemmas and their (presumed) semantic relationships to the original 210 words.

2.1 Chinese word list

We started this project by selecting a Chinese word list to be used in the pre-CWN. The selection criteria were that the list must be small in number but comprehensive in coverage, both in terms of distribution and semantic relations. In addition, since we would be testing the cross-lingual transportability of semantic relations, grammaticalized meanings that are language-dependent had to be avoided. Hence only nouns, verbs, adjectives, and adverbs are selected in this study. We took the 200 most frequent Chinese words as our starting set. Unfortunately, there were no adjectives among these 200 words.[2] To maintain semantic coverage, supplementary selection of the 10 most frequent adjectives from the corpus was made. After automatic extraction, additional manual adjustments were needed to increase compatibility with the TEDB. Since stative intransitive verbs (tagged VH) in the Sinica Corpus were as a rule translated as adjectives in English, we applied the adjectivization rule, i.e. X → X 的, to add the attributive particle "的" 'DE' to those words. This rendered the translation equivalences transparent and unambiguous, and established direct correspondence to English WN. This step increased precision when mapping Chinese words onto English synsets in the

[1] Although we have found instances of all WN semantic relations, we restrict our discussion in this paper to antonym, hypernym, and hyponym.

[2] In Chinese only modifiers that are exclusively attributive appear as adjectives. Modifiers that can be used predicatively take the form of stative intransitive verbs.

database. Eleven items were involved in this process. The complete list of 210 words (108 nouns, 41 verbs, 10 adjectives, and 51 adverbs) is given in APPENDIX I.

2.2 Translation equivalence database (TEDB)

The translation equivalence database was hand-crafted by the WordNet team at CKIP, Academia Sinica. First, all possible Chinese translations of an English synset word (from WN 1.5) are extracted from several available online bilingual (EC or CE) resources. These translation candidates were then checked by a team of translators with near-native bilingual ability. For each of the 99,642 English synsets, the translator selected the three most appropriate translation equivalents whenever possible. The translation equivalences were defaulted to lexicalized words, rather than descriptive phrases, whenever possible. The translation equivalences for 42,606 synsets were manually verified before the start of this study. Other synsets were either still being verified or failed to be translated as Chinese lexeme.

Our 210 PreCWN target lemmas were found as translation equivalences for 496 English lemmas.[3] The categorical distributions of these English lemmas are: 195 nouns, 161 verbs: 47 adjectives, and 94 adverbs. These 496 English lemmas belong to 441 WordNet synsets, since each synset contains one or more lemmas. Thus each Chinese lemma corresponds to 2.13 English synsets on average. If WordNet synsets are approximation word senses, then each Chinese lemma has 2.13 senses on average. The above 441 synsets contain 597 English lemmas, including the original 496 words obtained directly through translation equivalences. Extending from the 441 English synsets through semantic relations as marked by WordNet, there are 1,056 additional synsets. These English synsets are linked to 1,743 Chinese words in our TEDB. The bootstrapping process is diagrammed below. Note that we use perpendicular hollow arrows to represent bilingual processes, and horizontal filled arrowed to indicate monolingual processes.

[3] Note that we allow up to three best Chinese translation equivalences for each English lemma and that a Chinese lemma can be used for the translation equivalence of one or more English lemmas.

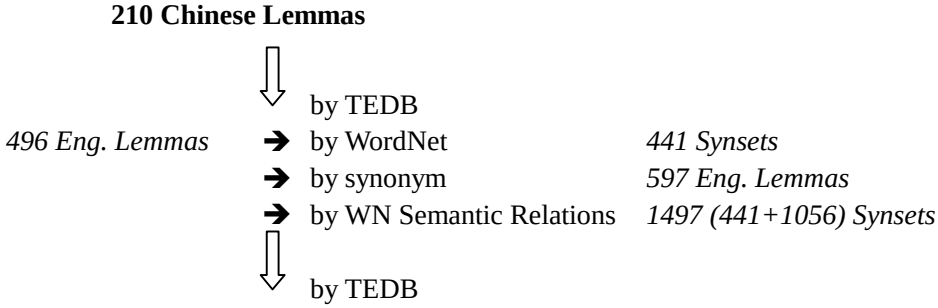

Diagram 1. Bootstrapping Chinese Semantic Relations from TEDB and English WordNet

In other words, based on TEDB and WordNet, the 210 original Chinese lemmas are now linked, through various candidate semantic relations, to 1,743 Chinese lemmas. On average, each Chinese lemma was linked with 8.3 inferred semantic relations. These numbers suggest that the attempt to bootstrap a Chinese wordnet from the English WN is promising, since it yields a sizable candidate set of semantically related lemmas for the original list.

2.3 Evaluation of the translation equivalences and inferred semantic relations

The automatic mapping described in the last section expanded the 210 Chinese lemmas to 1,743 lemmas bootstrapped through English semantic relations and TEDB. These bootstrapped semantic relations are manually evaluated by linguists. We first assumed the bootstrapped relations to be correct. That is, if A is the hypernym of B in English WN, we would expect A_C (the Chinese correspondence of A) is also the hypernym of B_C. Our initial evaluation classifies the inferred semantic relations into one of the following six categories: Correct, Incorrect, Other Relations, Revisable, Not Lexicalized, and Debatable. This tentative classification helps to facilitate the human verification process, since the linguists have not come up with a definite alternative when faced with the difficult cases of Debatable and Revisable.

No.	Label	Meaning
1	Debatable	assigned relation is hard to evaluate; needs further discussion
2	Correct	assigned relation is correct
3	Incorrect	assigned relation is wrong, and the two words have no other relations

4	Other Relation	assigned relation is wrong, but the two words have other semantic relations
5	Revisable	assigned relation is correct; the current Chinese translation is not a lexical word, but an appropriate word can be found
6	Not Lexicalized	assigned relation is correct; the Chinese translation is a phrase which cannot be lexicalized

After the first round, relations classified in the two difficult categories (i.e., Other and Revisable), are reëxamined and assigned to the other categories. The relations from the Not-Lexicalized category are discarded because we are only concerned with lexical semantic relations. Hence, in our study, only three categories are reported: Correct, Incorrect, and Other Relations.

Conceptually and procedure-wise, the evaluation consists of two parts: felicity of the translation equivalence, and validity of the semantic relation.

1) Felicity of the translation equivalence: Because linguistic ontologies vary, the way to describe a concept may not be exactly the same across languages. It is possible that a concept is lexicalized in one language but is represented by a phrase in another. Or, concepts lexicalized in two languages may only have an overlap in meaning but not be completely equivalent. Therefore, we have to determine whether our cross-lingual synonym pairs are actually semantically equivalent.

2) Validity of the inferred semantic relation: although we presumed the bootstrapped relations to be correct, we needed to verify our hypothesis. It is necessary to set up procedures for our human subjects to follow. The criteria that we apply were reported in Tsai, et al., (2002). The result of the evaluation process will be discussed in the following section.

3 Evaluation results and analysis

3.1 English-Chinese semantic relations

Huang, Tseng, & Tsai (2002) showed that semantic relations between a pair of bilingual translation equivalents are non-trivial and must be explicitly marked for language processing. This is because a pair of bilingual translation equivalents are not necessarily synonymous. A translation equivalent may have other semantic relations with the sense stipulated in the source-language word. Since our goal is to discover monolingual Chinese semantic relations through the bootstrapping from known monolingual English semantic relations, we need to know if the English-Chinese translation equivalents are synonymous. A priori, we expect the English semantic relations to be more reliably bootstrapped when there is a synonymous relation between

the two translation equivalents.[4] In addition, explicitly marking the semantic relations is also a crucial step in the evaluation of the quality and applicability of our English-Chinese TEDB.

As mentioned earlier, the 210 Chinese lemmas are used in translation of 496 English lemmas. In other words, they are involved in 496 E-C equivalent pairs. Table 1 shows the results of our evaluation of these pairs. The evaluation is on whether the English-Chinese pairs are actually synonymous, a null hypothesis relation of any translation correspondences. Our evaluation shows that 77% of the translation equivalences are also synonymous. In particular, verbs are least likely to be synonymous after translation; 30% of verbs have no synonymous relation with their translation. Note that the categories reported are English categories.

	Correct		Incorrect		Other Relation		Total	
Nouns	148	75.9%	33	17%	14	7.2%	195	100%
Verbs	112	70%	29	18.1%	19	11.9%	160	100%
Adjectives	39	83%	8	17.%	0	0%	47	100%
Adverbs	83	88.3%	8	8.5%	3	3.2%	94	100%
Total	382	77%	78	15.7%	36	7.3%	496	100%

Table 1: C-Word to E-Word Equivalences (Total Pairs=496)

To be exhaustive in locating all possible semantic relations, we expanded the list of equivalence pairs by including all WordNet synonyms, as defined by all lemmas from the 441 synsets covering the original 496 English lemmas. One hundred and one words are thus added, and the 597 bilingual pairs are evaluated in Table 2. In theory, adding monolingual synonyms should not affect the accuracy of translation equivalency. However, empirically, we know that some 'synonyms' are more equal than others. Hence we did get a slight drop in the percentage of bilingual synonymous relations, as expected. Please note that non-lexicalized (e.g., phrasal) correspondences are excluded, hence the numbers do not add up to 100%.

[4] On the other hand, if there is another definite semantic relation between the translation equivalents, we can conceivably compute that relation and incorporate it into the bilingual bootstrapping process. Huang, et al., (2002) discussed this possibility.

	Correct	Incorrect	Other Rel.
Nouns	54.6%	24.4%	12%
Verbs	57.1%	25.3%	9.9%
Adjectives	64.9%	14.0%	3.5%
Adverbs	79.9%	13.4%	2.7%
Total	62.7%	20.9%	8.2%

Table 2: Translation Equivalences with Synonym Expansion (Total Pairs=597)

The next step in bootstrapping to locate additional semantically related Chinese lemmas is to take advantage of all semantic relations encoded on the selected English translation equivalences. Note that the total yield at this stage is 1,743 English-Chinese pairs. We concentrate on three more easily definable (and more frequent) semantic relations: antonym (ANT); hypernym (HYP); and hyponym (HPO). Table 3 and Table 4 list the evaluation results based on the English categories of nouns and verbs.

Nouns

	Synonym	Incorrect	Other Relation	Others	Total
ANT	7	3	0	2	12
	58.3%	25%	0%	16.7%	100%
HYP	117	33	15	20	185
	63.2%	17.8%	8.1%	10.8%	100%
HPO	284	119	66	256	725
	39.2%	16.4%	9.1%	35.3%	100%
Total	408	155	81	278	922
	44.3%	16.8%	8.8%	30.2%	100%

Table 3: TE expanded with SR—Nouns (Total Pairs=922)

Verbs

	Synonym	Incorrect	Other Relation	Others	Total
ANT	8	6	0	9	23
	34.8%	26.1%	0%	39.1%	100%
HYP	61	18	6	2	87
	70.1%	20.7%	6.9%	2.3%	100%
HPO	118	81	19	74	292
	40.4%	27.7%	6.5%	25.3%	100%
Total	187	105	25	85	402
	46.5%	26.1%	6.2%	21.1%	100%

Table 4: TE expanded with SR—Verbs (Total Pairs=402)

Two observations can be immediately made involving the above data. The first is that once non-equivalency relationships are introduced, the default synonymy between each translation pair becomes even harder to maintain. The second is that the nature of the semantic relations does affect the translation synonymy. We shall elaborate on the second fact later. The data for adjectives and adverbs are more straightforward, since only antonyms are relevant.

Adjectives

	Correct	Incorrect	Other Rel.	Others	Total
ANTL	3	2	0	2	7
	42.9%	28.6%	0%	28.5%	100%

Table 5: TE expanded with SR—Adjectives (Total Pairs =7)

Adverbs

	Correct	Incorrect	Other Rel.	Others	Total
ANT	7	1	0	2	10
	70%	10%	0%	20%	100%

Table 6: TE expanded with SR—Adverbs (Total Pairs=10)

Note that adjectives and adverbs offer a very low number of inferable semantic relations (7 and 10 respectively, see Tables 5 and 6). Due to this sparseness of data, we shall not make any general claims regarding either adjectives or adverbs.

It has been suggested that high frequency words are more likely to be polysemous (Ahrens 1999). With the categorical-ambiguity data from the Sinica Corpus, Huang, Chen, & Shen (2002) show that this tendency holds in Chinese, with the word's category being another important factor. In English, this tendency is instantiated as the high number of senses for frequent words, such as the 26 senses of "make" and the 23 senses of "take" in WordNet. Recall that the 210 original lemmas in Chinese are the most frequent words in each category. In other words, the additional lemmas linked through inferred semantic relations are 1) less frequent, and 2) in many cases, alternative translations (compare the most frequent and typical one from the list of 210) for the English lemma. Based on the two above observations, it is reasonable to expect that the immediate pairs based on the 210 lemmas (Tables 1 and 2) are more synonymous than the larger set including the inferred pairs (Tables 4-6).

To improve the accuracy of the TEDB, as well as to investigate the conditions under which non-synonyms are most likely to be picked as translation equivalents, we looked at all instances where the bilingual synonymous relations did not hold. The

results are summarized in APPENDIX II. Referring to the chart, we can make two generalizations. First, the more polysemous words are more likely to lead to non-synonymous translation equivalents. Second, more abstract meanings also may lead to non-synonymous translations. The first generalization is well attested in our data, where the most polysemous nouns (e.g., 工作 gong1zuo4 'job, work, etc.' and 人 ren2 'human, -er, man, etc.') and the polysemous verbs (e.g., 成爲 cheng2wei2 'become, form, change into, etc.' and 進行 jin4xing2 'proceed, perform, undertake, etc.') all have non-synonymous translation equivalents. There are two possibilities: The first and more intuitive one is that, the more complicated a group of polysemous senses are, the less likely it is that they will be classified identically across different languages. The second possible explanation is that the number of completely synonymous translation equivalences across different languages is simply a function of the number of senses involved. The more senses a word-form has, the more likely it is that some of them will be translated with a non-synonymous word. These two explanations are not necessarily exclusive of each other. But we need more in-depth study to tease the actual relations out. The second observation seems true, but needs to be supported by objective data.

As for the second generalization suggesting that abstract meanings may lead to non-synonymous translation, this can be demonstrated by the examples:方面 fang1mian4 'side, facet, etc.' and 系統 xi4tong3 'system, lineage, etc.' in nouns; 認爲 ren4wei2 'think, seem, consider, etc.' and 讓 rang4 'let, allow, give, etc.' in verbs. Many instances involving the abstract senses can be improved when a suitable ontology or conceptual hierarchy is introduced to disambiguate. Since these examples all happen to be highly polysemous, it is likely that abstractness will not be a direct contributing factor in non-synonymous translation equivalency.

In sum, we conclude from the above evaluation that the more polysemous a Chinese lemma is, the least likely that there would be an exact set of matching English synsets. On the other hand, if only direct translation equivalences are considered, 77% are bilingual synonyms. Since the data is compiled on 210 highly polysemous lemmas, we expect that the rate of bilingual synonymy will be higher for other less frequent lemmas. These findings suggest that the null hypothesis that translation equivalence pairs are synonymous is reasonable for less frequent words, while questionable for frequent words. Hence the use of TEDB in bilingual language processing must be sensitive to lexical frequency.

3.2 Semantic relation of Chinese-Chinese pairs

In this section, we examine the cross-lingual inferability of semantic relations. Since the cross-linguistic definition of non-synonymous semantic relations requires further

clarification to establish principled tests, we limit our examination to the Chinese lemmas that are both a translation equivalent of an English WN entry and are considered to have the synonymous semantic relation to that entry (cf. Huang, Tseng, & Tsai 2002). This will also ensure that the cross-lingual inference of the semantic relation is not distorted by the translation process.

From the 148 nouns where the English and Chinese translation equivalents are also synonymous, there are 357 pairs of semantic relations that are marked in the English WN and are therefore candidates for inferred relations in Chinese. The precision of the inferred semantic relations is tabulated below.

	Correct		Others		Total	
ANT	8	100%	0	0%	8	100%
HYP	70	79.5%	18	20.5%	88	100%
HPO	238	91.2%	23	8.8%	261	100%
Total	316	88.5%	41	11.5%	357	100%

Table 7: Precision of English-to-Chinese SR Inference (Nouns)

Note that since the evaluative tags of “revisable” and “not lexicalized” refer specifically to translation equivalency, not the semantic relation, they are not included for the current study. The figures show that when a direct synonymous relation can be established between the translation equivalence pairs (which is about 44.25% of the noun translation equivalents we studied), up to 90% precision can be achieved when we directly transport English WN semantic relations to Chinese. And among the other semantic relations examined, the antonymous relation is the most reliably transportable cross-linguistically.

From the 112 verbs where the English and Chinese translation equivalents are also synonymous, there are 155 pairs of semantic relations that are marked in the English WN and are therefore candidates for inferred relations in Chinese. The precision of the inferred semantic relations is tabulated below.

	Correct		Incorrect		Total	
ANT	14	100%	0	0%	14	100%
HYP	35	70%	15	30%	50	100%
HPO	75	82.4%	16	17.6%	91	100%
Total	124	80%	31	20%	155	100%

Table 8: Precision of English-to-Chinese SR Inference (Verbs)

Similar to the results for nouns, the antonymous relation appears reliable in verbs as well. As to the other types of relations, the correctness rates seem to be slightly lower than nouns. The precision for English-to-Chinese semantic relation inference is 80% for verbs.

The observed discrepancy in terms of semantic relation inference between nouns and verbs deserves in-depth examination. First, the precision of inference in nouns is 8.52% higher than in verbs, which merits closer examination. Second, the contrast may not be attributed to a specific semantic relation. Both nouns and verbs have similar precision patterns for the three semantic relations that we studied. Antonymy inference is highly reliable in both categories (both 100%). Hyponymous inference comes second, about 12% higher than hypernymous inference in each category (the difference is 11.64% for nouns and 12.42% for verbs). And last but not least, the precision gaps between nouns and verbs, when applicable, are similar for different semantic relations (9.55% for hypernyms and 8.77% for hyponyms). All the above facts support the generalization that noun semantic relations are more reliably inferred across languages than verb semantic relations. A plausible explanation for this generalization is the difference in mutability of noun and verb meaning, as reported in Ahrens (1999). Ahrens demonstrated with off-line psycholinguistic experiments that verb meanings are more mutable than noun meanings. She also reported that verb meaning has the tendency to change in coërcive contexts. We may assume that making the cross-lingual transfer is a coërcive context in terms of sense identification. Taking the mutability account, we can predict that since verb meanings are more likely than nouns to change given coërcive conditions, the changes will affect their semantic relations. Hence the precision for semantic relation inference is lower for verbs than for nouns.

In the above discussion, we observed that the three semantic relations seem to offer clear generalizations with regard to the precision of inferences. This observation can be highlighted when the evaluation results are represented according to semantic relations without specifying categories, as in Table 9.

	Correct		Incorrect		Total	
ANT	22	100%	0	0%	22	100%
HYP	105	76.1%	33	13.9%	138	100%
HPO	313	88.9%	39	11.1%	352	100%
Total	440	85.9%	72	14.1%	512	100%

Table 9: Combined Precision of English-to-Chinese SR Inference (Nouns+Verbs)

Two generalizations emerge from the data above and call for explanation: First, the inference of antonymous relations is highly reliable; second, the inference of hypernymous relations is more reliable than the inference of hyponymous relations.

The fact that English-to-Chinese inference of antonymous relations is highly precise may be due to either of the following facts. On one hand, since the number of antonymous relations encoded is relatively low (only 22 altogether), these lemmas may well be the most typical instances of their respective synsets. In other words, there is no room for variation. On the other hand, we observe that a pair of antonyms can be extremely close in meaning and differ in only one feature. In other words, an antonym (of any word) is simply a privileged (near) synonym whose meaning offers contrast in one particular semantic dimension. Since antonymy presupposes synonymous relations, it preserves the premise of our current semantic relation inference, leading to the high inference precision.

The fact that hyponymous relations can be more reliably inferred cross-linguistically than hypernymous relations is somewhat surprising, since these are symmetric semantic relations within a language. That is, if A is a hypernym of B, then B is a hyponym of A. Logically, there does not seem to be any reason for the two relations to have disjoint distribution when transplanted into another language. In other words, a naïve model would predict that both relations would be equally successful. However, if we study the conceptual nature of the semantic relations more carefully, there seems to be a plausible explanation.

We noted before that our cross-lingual inference of semantic relations is based on the synonymous relations between the translation equivalents. This fact actually leads to very different lexical entailments for inferred hyponym and hypernym relations. First, the stipulation that an English word Eng_1 has a word W_i as a hyponym HPO_1 entails that:

> Eng_1 lexically represents a conceptual class Con_1, such that the meaning of W_i ISA Con_1.

Second, on the other hand, the stipulation that an English word Eng_1 has a word W_i as a hypernym HYP_1 does not lexically entail such an equivalent class, since that class is not lexically referred to. It entails that:

> There is a non-specified conceptual class Con_e, such that the meaning of Eng_1 ISA Con_e.

Since our inference is based on the synonymous relation of the Chinese translation equivalent to the English word Eng_1, we can assume that the conceptual class Con_1 represented by that word is largely preserved. Hence the inference of hyponym relations has high precision. Inference based on hypernym relations, however, has no such lexically specified conceptual foundation to rely on. Failure in inference can in most

cases be attributed to the fact that the intended hypernym class has no synonymous translation equivalent in Chinese. In other words, the success of inference of the hypernymous relation must presuppose an additional semantic condition. Hence its lower precision can be expected.

To sum up, our preliminary evaluation found that the precision of cross-lingual inference of semantic relations can be higher than 90% if the inference does not require other conceptual/semantic relations other than the synonymy of the translation equivalents. On the other hand, an additional semantic relation, such as the equivalence of the hypernym node in both languages when inferring hyponym relations, seems to bring the precision rate down by about 10%.

4. Conclusion and further work

In this study, we first mapped our input lemmas to the TEDB in order to figure out their corresponding English synsets. By assuming the bootstrapped semantic relations from WN would hold true for Chinese, we allowed the automatic linking between an input lemma and its semantically related words. Then, we manually checked each link. We found that the database provides moderate results, but we also proposed that the highly frequent words chosen could have slightly worsened the results due to their inherent polysemy. Second, we analyzed the equivalence pairs with non-synonymous features in TEDB. We suggested that polysemous words and words with abstract meanings tend not to have exact equivalents in English. Last, we evaluated the semantic relations in Chinese inherited from their equivalent translations. It showed that once the translation is equivalent, the automatically assigned relation in Chinese turned out to be correct with a very high probability.

Since higher frequency words tend to be more polysemous, our current study of the highest frequency words should in theory return lower-bound results. Thus, even though the current result is only fair, it would not be naïve to expect that better results can be obtained with less frequent words. We are currently working on another CWN sample of input words of medium frequencies. Using this bottom-up strategy, we aim to construct a Chinese wordnet and obtain generalizations on criteria for Chinese sense distinction at the same time.

Other related work already completed includes a paper on the theoretical bases and implications of the bootstrapping procedure (Huang, Tseng, & Tsai 2002), and a pilot study on enriching domain information through automated linking to a Chinese machine-readable dictionary (Chang, et al., 2002). There is also on-going work on developing consistent criteria and working practices for establishing sense distinctions in Chinese.

Appendix I:

Lists of All Chinese Lemmas in the PreCWN (with Sinica Corpus Frequency and POS)

Nouns

NO.	WORD	POS	Freq.	NO.	WORD	POS	Freq.	NO.	WORD	POS	Freq.
1	一	Neu	58388	37	目前	Nd	4867	73	元	Nf	3132
2	個	Nf	41077	38	中	Ncd	4828	74	網路	Na	3093
3	我	Nh	40332	39	工作	Na	4620	75	日本	Nc	3061
4	這	Nep	33659	40	全	Neqa	4569	76	中心	Nc	3041
5	他	Nh	30025	41	這些	Neqa	4391	77	地方	Na	2990
6	人	Na	24269	42	裡	Ncd	4293	78	關係	Na	2951
7	我們	Nh	18152	43	現在	Nd	4236	79	市場	Nc	2950
8	你	Nh	17298	44	時候	Na	4179	80	前	Ng	2944
9	種	Nf	12263	45	時間	Na	4044	81	老師	Na	2871
10	中	Ng	12231	46	事	Na	4008	82	學校	Nc	2857
11	她	Nh	10776	47	中國	Nc	3900	83	經濟	Na	2831
12	那	Nep	10740	48	第一	Neu	3879	84	其他	Neqa	2818
13	上	Ncd	10619	49	美國	Nc	3826	85	家	Nc	2813
14	年	Nf	10127	50	幾	Neu	3721	86	教育	Na	2778
15	時	Ng	9565	51	系統	Na	3631	87	裡	Ng	2704
16	自己	Nh	9069	52	政府	Na	3612	88	方面	Na	2658
17	他們	Nh	8818	53	大家	Nh	3565	89	很多	Neqa	2640
18	兩	Neu	8692	54	國家	Na	3550	90	同時	Nd	2640
19	各	Nes	8651	55	許多	Neqa	3548	91	電腦	Na	2621
20	上	Ng	8650	56	生活	Na	3542	92	心	Na	2606
21	後	Ng	7752	57	大學	Nc	3508	93	企業	Na	2588
22	者	Na	7221	58	研究	Na	3485	94	臺灣	Nc	2572
23	每	Nes	7207	59	本	Nes	3462	95	空間	Na	2553
24	次	Nf	7087	60	二	Neu	3452	96	五	Neu	2546
25	三	Neu	6954	61	活動	Na	3432	97	國內	Nc	2536
26	什麼	Nep	6729	62	該	Nes	3380	98	今天	Nd	2536
27	問題	Na	6683	63	世界	Nc	3375	99	們	Na	2523
28	其	Nep	6667	64	四	Neu	3367	100	之後	Ng	2495
29	此	Nep	6599	65	方式	Na	3362	101	人員	Na	2486
30	台灣	Nc	6414	66	內	Ncd	3354	102	產品	Na	2457
31	位	Nf	6015	67	項	Nf	3328	103	資料	Na	2449
32	學生	Na	5523	68	下	Ng	3299	104	資訊	Na	2443
33	公司	Nc	5421	69	環境	Na	3276	105	先生	Na	2423
34	社會	Na	5282	70	一些	Neqa	3238	106	地	Na	2419
35	天	Nf	5038	71	文化	Na	3216	107	未來	Nd	2370
36	它	Nh	4964	72	孩子	Na	3201	108	大陸	Nc	2358

Verbs

NO.	WORD	POS	Freq.	NO.	WORD	POS	Freq.	NO.	WORD	POS	Freq.
1	是	SHI	84014	19	使	VL	4645	37	進行	VC	2963
2	有	V_2	45823	20	覺得	VK	4440	38	提供	VD	2869
3	說	VE	19625	21	使用	VC	4415	39	指出	VE	2836
4	大	VH	11577	22	知道	VK	4160	40	發展	VC	2796
5	大的			23	這樣	VH	4138	41	成爲	VG	2774
6	爲	VG	8369	24	這樣的			42	多	VH	2751
7	好	VH	8304	25	認爲	VE	4070	43	多的		
8	好的			26	到	VCL	3739	44	吃	VC	2636
9	讓	VL	6624	27	希望	VK	3365	45	發現	VE	2626
10	做	VC	6597	28	高	VH	3207	46	一樣	VH	2619
11	沒有	VJ	6510	29	高的			47	一樣的		
12	想	VE	5898	30	不同	VH	3113	48	服務	VC	2573
13	表示	VE	5504	31	不同的			49	看到	VE	2551
14	看	VC	5198	32	來	VA	3040	50	無	VJ	2519
15	小	VH	5051	33	對	VH	3038	51	開始	VL	2366
16	小的			34	對的			52	需要	VK	2350
17	新	VH	4978	35	重要	VH	2990				
18	新的			36	重要的						

Adjectives

NO.	WORD	POS	Freq.	NO.	WORD	POS	Freq.
1	主要	A	1757	11	真正	A	623
2	主要的			12	真正的		
3	一般	A	1491	13	唯一	A	570
4	一般的			14	唯一的		
5	共同	A	1253	15	最佳	A	465
6	共同的			16	最佳的		
7	基本	A	981	17	非	A	443
8	基本的			18	非的		
9	公共	A	785	19	國立	A	346
10	公共的			20	國立的		

Adverbs

NO.	WORD	POS	Freq.	NO.	WORD	POS	Freq.	NO.	WORD	POS	Freq.
1	不	D	39014	18	將	D	7858	35	不能	D	3145
2	了	Di	31873	19	更	D	7298	36	仍	D	3097
3	也	D	29646	20	才	Da	7266	37	太	Dfa	2893
4	就	D	29211	21	已	D	7256	38	應該	D	2839

5	都	D	20403	22	再	D	6563	39	非常	Dfa	2737
6	要	D	15955	23	只	Da	6521	40	便	D	2723
7	會	D	14066	24	則	D	6476	41	然後	D	2634
8	很	Dfa	13013	25	卻	D	6388	42	未	D	2629
9	能	D	11125	26	去	D	5748	43	無法	D	2591
10	著	Di	11026	27	並	D	4238	44	較	Dfa	2573
11	還	D	9698	28	一	D	4070	45	正	D	2573
12	可以	D	9671	29	過	Di	3945	46	不會	D	2573
13	最	Dfa	9416	30	可能	D	3928	47	曾	D	2558
14	來	D	8992	31	已經	D	3518	48	如何	D	2543
15	所	D	8873	32	應	D	3370	49	先	D	2465
16	可	D	8508	33	必須	D	3231	50	比較	Dfa	2426
17	又	D	8037	34	沒有	D	3195	51	在	D	2340

Appendix II:

The Correctness of Bilingual Synonymous Relations

Nouns

NO.	WORD	ES[5]	Correct	percentage	Incorrect	percentage	Other Rel.	percentage
1	工作	11	6	54.55%	2	18.18%	3	27.27%
2	人	10	6	60%	2	20%	2	20%
3	教育	7	4	57.14%	1	14.29%	2	28.57%
4	公司	5	2	40%	2	40%	1	20%
5	大學	5	4	80%	1	20%	0	0%
6	問題	5	5	100%	0	0%	0	0%
7	第一	5	5	100%	0	0%	0	0%
8	關係	5	4	80%	0	0%	1	20%
9	市場	4	3	75%	1	25%	0	0%
10	活動	4	3	75%	1	25%	0	0%
11	環境	4	3	75%	1	25%	0	0%
12	天	4	4	100%	0	0%	0	0%
13	時間	4	4	100%	0	0%	0	0%
14	現在	4	4	100%	0	0%	0	0%
15	先生	4	0	0%	0	0%	4	100%
16	元	3	1	33.33%	2	66.67%	0	0%
17	方面	3	1	33.33%	2	66.67%	0	0%
18	各	3	2	66.67%	1	33.33%	0	0%
19	幾	3	2	66.67%	1	33.33%	0	0%
20	學校	3	2	66.67%	1	33.33%	0	0%

[5] The number of translationally corresponding English Synsets.

21	中心	3	3	100%	0	0%	0	0%
22	方式	3	3	100%	0	0%	0	0%
23	世界	3	3	100%	0	0%	0	0%
24	前	3	3	100%	0	0%	0	0%
25	研究	3	3	100%	0	0%	0	0%
26	資料	3	3	100%	0	0%	0	0%
27	系統	2	0	0%	2	100%	0	0%
28	政府	2	0	0%	2	100%	0	0%
29	企業	2	1	50%	1	50%	0	0%
30	許多	2	1	50%	1	50%	0	0%
31	一	2	2	100%	0	0%	0	0%
32	之後	2	2	100%	0	0%	0	0%
33	文化	2	2	100%	0	0%	0	0%
34	目前	2	2	100%	0	0%	0	0%
35	全	2	2	100%	0	0%	0	0%
36	地方	2	2	100%	0	0%	0	0%
37	年	2	2	100%	0	0%	0	0%
38	次	2	2	100%	0	0%	0	0%
39	自己	2	2	100%	0	0%	0	0%
40	其他	2	2	100%	0	0%	0	0%
41	社會	2	2	100%	0	0%	0	0%
42	空間	2	2	100%	0	0%	0	0%
43	家	2	2	100%	0	0%	0	0%
44	項	2	2	100%	0	0%	0	0%
45	經濟	2	2	100%	0	0%	0	0%
46	資訊	2	2	100%	0	0%	0	0%
47	種	2	2	100%	0	0%	0	0%
48	時候	2	1	50%	0	0%	1	50%
49	下	1	0	0%	1	100%	0	0%
50	中國	1	0	0%	1	100%	0	0%
51	生活	1	0	0%	1	100%	0	0%
52	後	1	0	0%	1	100%	0	0%
53	國家	1	0	0%	1	100%	0	0%
54	產品	1	0	0%	1	100%	0	0%
56	網路	1	0	0%	1	100%	0	0%
57	學生	1	0	0%	1	100%	0	0%
58	一些	1	1	100%	0	0%	0	0%
59	二	1	1	100%	0	0%	0	0%
60	人員	1	1	100%	0	0%	0	0%
61	三	1	1	100%	0	0%	0	0%

62	大陸	1	1	100%	0	0%	0	0%
63	五	1	1	100%	0	0%	0	0%
64	今天	1	1	100%	0	0%	0	0%
65	內	1	1	100%	0	0%	0	0%
66	心	1	1	100%	0	0%	0	0%
67	日本	1	1	100%	0	0%	0	0%
68	台灣	1	1	100%	0	0%	0	0%
69	四	1	1	100%	0	0%	0	0%
70	本	1	1	100%	0	0%	0	0%
71	未來	1	1	100%	0	0%	0	0%
72	同時	1	1	100%	0	0%	0	0%
73	地	1	1	100%	0	0%	0	0%
74	老師	1	1	100%	0	0%	0	0%
75	每	1	1	100%	0	0%	0	0%
76	事	1	1	100%	0	0%	0	0%
77	兩	1	1	100%	0	0%	0	0%
78	其	1	1	100%	0	0%	0	0%
79	者	1	1	100%	0	0%	0	0%
80	孩子	1	1	100%	0	0%	0	0%
81	很多	1	1	100%	0	0%	0	0%
82	國內	1	1	100%	0	0%	0	0%
83	電腦	1	1	100%	0	0%	0	0%
84	臺灣	1	1	100%	0	0%	0	0%

Verbs

NO.	WORD	ES	Correct	percentage	Incorrect	percentage	Other Rel.	percentage
1	使	10	5	50%	2	20%	3	30%
2	重要	10	7	70%	2	20%	1	10%
3	開始	9	4	44.44%	2	22.22%	3	33.33%
4	不同	9	6	66.67%	2	22.22%	1	11.11%
5	成爲	7	3	42.86%	4	57.14%	0	0%
6	進行	7	4	57.14%	3	42.86%	0	0%
7	知道	7	5	71.43%	2	28.57%	0	0%
8	發現	7	5	71.43%	0	0%	2	28.57%
9	是	6	3	50%	3	50%	0	0%
10	認爲	6	3	50%	3	50%	0	0%
11	需要	5	4	80%	1	20%	0	0%
12	做	5	3	60%	0	0%	2	40%
13	說	5	3	60%	0	0%	2	40%
14	讓	5	3	60%	0	0%	2	40%

15	提供	5	5	100%	0	0%	0	0%
16	吃	4	3	75%	0	0%	1	25%
17	有	4	3	75%	0	0%	1	25%
18	小	4	4	100%	0	0%	0	0%
19	服務	3	1	33.33%	1	33.33%	1	33.33%
20	表示	3	2	66.67%	1	33.33%	0	0%
21	對	3	2	66.67%	1	33.33%	0	0%
22	大	3	3	100%	0	0%	0	0%
23	好	3	3	100%	0	0%	0	0%
24	發展	3	3	100%	0	0%	0	0%
25	覺得	3	3	100%	0	0%	0	0%
26	無	2	1	50%	1	50%	0	0%
27	多	2	2	100%	0	0%	0	0%
28	希望	2	2	100%	0	0%	0	0%
29	沒有	2	2	100%	0	0%	0	0%
30	看	2	2	100%	0	0%	0	0%
31	高	2	2	100%	0	0%	0	0%
32	想	2	2	100%	0	0%	0	0%
33	新	2	2	100%	0	0%	0	0%
34	使用	1	0	0%	1	100%	0	0%
35	一樣	1	1	100%	0	0%	0	0%
36	來	1	1	100%	0	0%	0	0%
37	到	1	1	100%	0	0%	0	0%
38	指出	1	1	100%	0	0%	0	0%
39	爲	1	1	100%	0	0%	0	0%
40	看到	1	1	100%	0	0%	0	0%
41	這樣	1	1	100%	0	0%	0	0%

Adjectives

NO.	WORD	ES	Correct	percentage	Incorrect	percentage	Other Rel.	percentage
1	主要	11	8	72.73%	3	27.27%	0	0%
2	一般	11	10	90.91%	1	9.09%	0	0%
3	真正	7	6	85.71%	1	14.29%	0	0%
4	基本	5	3	60%	2	40%	0	0%
5	唯一	4	4	100%	0	0%	0	0%
6	共同	3	2	66.67%	1	33.33%	0	0%
7	公共	2	2	100%	0	0%	0	0%
8	最佳	2	2	100%	0	0%	0	0%
9	非	1	1	100%	0	0%	0	0%
10	國立	1	1	100%	0	0%	0	0%

Adverbs

NO.	WORD	ES	Correct	percentage	Incorrect	percentage	Other Rel.	percentage
1	非常	11	7	63.64%	2	18.18%	2	18.18%
2	要	6	5	83.33%	1	16.67%	0	0%
3	會	4	1	25%	3	75%	0	0%
4	很	4	2	50%	1	25%	1	25%
5	可以	4	4	100%	0	0%	0	0%
6	不	3	3	100%	0	0%	0	0%
7	太	3	3	100%	0	0%	0	0%
8	無法	3	3	100%	0	0%	0	0%
9	應該	3	3	100%	0	0%	0	0%
10	比較	2	1	50%	1	50%	0	0%
11	又	2	2	100%	0	0%	0	0%
12	也	2	2	100%	0	0%	0	0%
13	不能	2	2	100%	0	0%	0	0%
14	可能	2	2	100%	0	0%	0	0%
15	正	2	2	100%	0	0%	0	0%
16	先	2	2	100%	0	0%	0	0%
17	更	2	2	100%	0	0%	0	0%
18	沒有	2	2	100%	0	0%	0	0%
19	並	2	2	100%	0	0%	0	0%
20	則	2	2	100%	0	0%	0	0%
21	曾	2	2	100%	0	0%	0	0%
22	然後	2	2	100%	0	0%	0	0%
23	過	2	2	100%	0	0%	0	0%
24	應	2	2	100%	0	0%	0	0%
25	一	1	1	100%	0	0%	0	0%
26	已	1	1	100%	0	0%	0	0%
27	已經	1	1	100%	0	0%	0	0%
28	才	1	1	100%	0	0%	0	0%
29	不會	1	1	100%	0	0%	0	0%
30	仍	1	1	100%	0	0%	0	0%
31	可	1	1	100%	0	0%	0	0%
32	只	1	1	100%	0	0%	0	0%
33	必須	1	1	100%	0	0%	0	0%
34	未	1	1	100%	0	0%	0	0%
35	再	1	1	100%	0	0%	0	0%
36	在	1	1	100%	0	0%	0	0%
37	如何	1	1	100%	0	0%	0	0%
38	來	1	1	100%	0	0%	0	0%

39	便	1	1	100%	0	0%	0	0%
40	卻	1	1	100%	0	0%	0	0%
41	能	1	1	100%	0	0%	0	0%
42	將	1	1	100%	0	0%	0	0%
43	都	1	1	100%	0	0%	0	0%
44	最	1	1	100%	0	0%	0	0%
45	就	1	1	100%	0	0%	0	0%
46	較	1	1	100%	0	0%	0	0%
47	還	1	1	100%	0	0%	0	0%

References

Ahrens, Kathleen. 1999. The mutability of noun and verb meaning. *Chinese Language and Linguistics*, vol.5: *Interactions in Language*, ed. by Y. Yin, I. Yang and H. Chan, 335-371. Taipei: Academia Sinica.

Chang, Echa, Chu-Ren Huang, Sue-Jin Ker, and Chang-Hua Yang. 2002. Induction of classification from lexicon expansion: Assigning domain tags to WordNet entries. *Proceedings of the COLING2002 Workshop "SemaNet: Building and Using Semantic Networks"*, ed. by Grace Ngai, Pascale Fung, and Kenneth W. Church, 80-86.

Fellbaum, Christiane. (ed.) 1998. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

Huang, Chu-Ren, Chao-Ran Chen, and Claude C. C. Shen. 2002. The nature of categorical ambiguity and its implications for language processing: A corpus-based study of Mandarin Chinese. *Sentence Processing in East Asian Languages*, ed. by Mineharu Nakayama. Stanford: CSLI Publications.

Huang, Chu-Ren, E. I-Ju Tseng, and Dylan B. S. Tsai. 2002. Translating lexical semantic relations: The first step towards multilingual WordNets. *Proceedings of the COLING2002 Workshop "SemaNet: Building and Using Semantic Networks"*, ed. by Grace Ngai, Pascale Fung, and Kenneth W. Church, 2-8.

Pianta, Emanuel, L. Benitivogli, C. Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. *Proceedings of the 1st International WordNet Conference*, 293-302. Mysore, India.

Tsai, Dylan B. S., Chu-Ren Huang, Shu-chuan Tseng, Jen-yi Lin, Keh-jiann Chen, and Yuan-hsun Chuang. 2002. Criteria for defining and determining semantic relations in Mandarin Chinese. *Journal of Chinese Information Processing* 16.4:21-31. (In Chinese)

Vossen, P. (ed.) 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic.

[Received 5 September 2002; revised 20 May 2003; accepted 26 May 2003]

Chu-Ren Huang
Institute of Linguistics, Preparatory Office
Academia Sinica
130, Sec. 2, Academia Road
Taipei 115, Taiwan
churen@gate.sinica.edu.tw

跨語言語意關係轉換的適用性—利用英語詞網關係構建中文詞網之研究

黃居仁　曾意茹　蔡柏生　莫　非
中央研究院

多語詞網間對應關係的建立，是多語知識處理的基本要件之一。本文主張多語詞網間的對應關係應該建立在語意關係上，而不是在翻譯或知識本體上。我們將雙語間的對譯關係重新定義爲兩個語言間的詞彙語意關係。我們可藉著這些關係的邏輯推論來預測來源語的語意關係是否可在目標語中適用。以上的主張，我們用中文中的 210 最常用詞形進行實驗，主要用它們在英語詞網中對譯詞的語意關係作假設的跨語言語意關係轉換。結果顯示跨語言的詞彙語意關係，在符合邏輯推導關係的條件下，可以準確預測。此研究對將來藉助現有語言詞網，爲語言資源較貧乏的語言建立新詞網的研究，有正面的意涵。

關鍵詞：（詞彙）語意關係，詞網，雙語語意關係，翻譯對應，中文詞網