Cross-lingual Conversion of Lexical Semantic Relations: Building Parallel Wordnets

Chu-Ren Huang[1] , I-Li Su[1], Jia-Fei Hong[1], Xiang-Bing Li[2]
[1]. Institute of Linguistics
[2]. Institute of Information Science
Academia Sinica,
No.128 Academic Sinica Road, SEC.2 Nankang,
Taipei 115, Taiwan
[1].{ churen, isu, jiafei }@gate.sinica.edu.tw

Abstract

Parallel wordnets built upon correspondences between different languages can play a crucial role in multilingual knowledge processing. Since there is no homomorphism between pairs of monolingual wordnets, we must rely on lexical semantic relation (LSR) mappings to ensure conceptual cohesion. In this paper, we propose and implement a model for bootstrapping parallel wordnets based on one monolingual wordnet and a set of cross-lingual lexical semantic relations. In particular, we propose a set of inference rules to predict Chinese wordnet structure based on English wordnet and English-Chinese translation relations. We show that this model of parallel wordnet building is effective and achieves higher precision in LSR prediction.

1 Introduction

A knowledgebase which systemizes lexical and conceptual information of human knowledge is a basic infrastructure for Natural Language Processing (NLP) applications. Wordnets, pioneered by the Princeton WordNet (WN, Fellbaum 1998), and greatly enriched by EuroWordnet (EWN, Vossen 1998), have become the standard for a lexical knowledgebase enriched with lexical semantic relations. In addition to the multilingual architecture of EWN, there are some proposals to construct the expansion for monolingual wordnets to parallel wordnet systems, such as Pianta and Girardi (2002). However, the construction of multilingual wordnets eventually faces the challenge of low-density languages, which is dealt with in Huang, et al. (2002). Low-density languages, as opposed to high-density languages, usually refer to languages that are not spoken by a large number of people. However, there is neither a direct correspondence between language population and language technology, nor an objective population number that defines density level. In this work, we use the availability of language resources instead to define language density. That is, low-density languages are languages that do not have enough language resources to support fully automated language processing, such as machine translation. In our current line of work, we (Huang et al. 2002) refer to low-density languages as those which do not have enough existing resources for semi-automatic construction of monolingual wordnet.

There are two alternative approaches to build parallel wordnets, as shown in Figure 1. The first approach relies on two fully annotated monolingual wordnets with synsets and LSR's. The second approach requires only one fully annotated WN in addition to LSR-based cross-lingual translation correspondences.

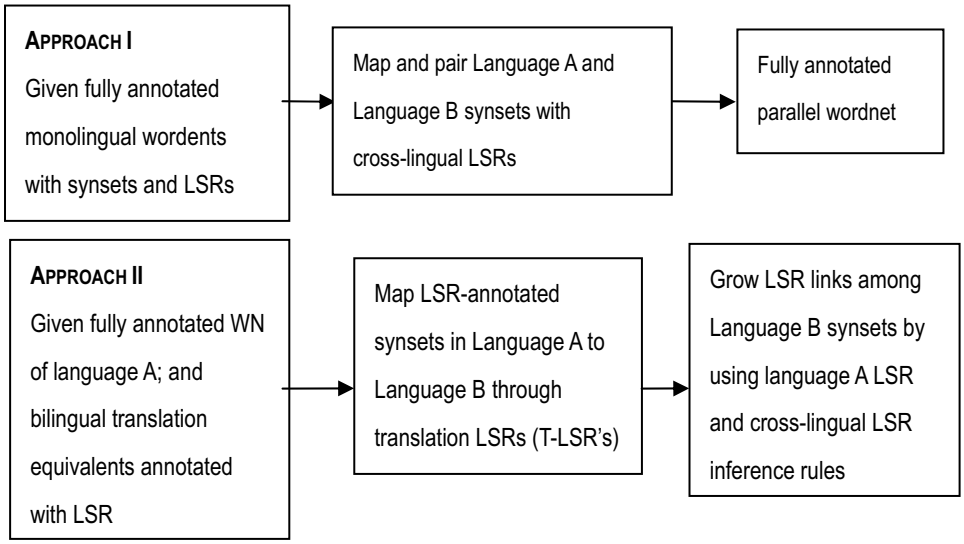

Figure1. Two Approaches to Building Bilingual Wordnets

Approach I maps and pairs Language A synsets with Language B synsets and annotates cross-lingual LSR's. The result is a fully annotated parallel wordnet. Approach II maps language A synsets to language B through translation equivalents. After language B synsets are thus established, language B LSR's are predicted based on corresponding LSR's in language A. A new set of monolingual LSR's is bootstrapped and predicted basing on inference rules governed by translation LSR's (T-LSR's). In general, approach I applies to high-density languages while approach II applies to low-density languages. In this paper, we will focus on the application of approach II to build a Chinese Wordnet with conceptual cohesion.

The current model was first explored in Huang et al. (2003). This previous study covered 210 lemmas, consisted of the top ranked lemmas in each part-of-speech (POS). The translation LSR's discussed in the previous model were antonymy, hypernymy and hyponymy. In this current work, we expand our study to all possible LSR's as well as to all the bilingual lexical pairs in our English-Chinese translation equivalents databases. Moreover, the LSR's in Princeton WordNet are again used as the basis for bootstrapping. In addition, we establish a set of evaluation for the results. The approach will be evaluated in term of both the precision of prediction and the confidence of prediction. We aim to show that T-LSR's bootstrapped approach does provide an effective model for building parallel wordnets for low-density languages.

After the introduction, the main part of this paper consists of the following sections: in section 2, we briefly introduce the existing resources required for this work. We discuss methodology of T-LSR bootstrapping step by step in section 3. A series of LSR-predicting inference rules are also given in this section. In section 4, we plan to evaluate the results of our experiment and demonstrate the feasibility of maintaining conceptual cohesion in cross-lingual LSR mapping.

2 Required Resources: ECTED and WN

As we mentioned above, the T-LSR approach to parallel wordnet requires two language resources: a fully annotated monolingual wordnet and a set of translation LSR's to map the wordnet information to the target language. In our current study, we use the English WN as the source of synset and LSR information. The semantic relation between an English synset and its Chinese translation is based on The English-Chinese Translation Equivalents Database (ECTED, Huang et al. 2002).

2.1 The English-Chinese Translation Equivalents Databases (ECTED)

The basic idea of ECTED is to provide the Chinese translation equivalents for each

WN English synset. Our ECTED was bootstrapped with a combined lexical knowledgebase integrating at least four English-Chinese or Chinese-English bilingual resources. Based on this combined LKB, a group of translators chose (or created) up to three best translation equivalents for each WN synset. In addition, for each English-Chinese translation equivalent, a lexical semantic relation is annotated. In addition to synonym, the semantic relations marked including antonym, hypernym, hyponym, holonym, meronym, and near-synonym. We use all semantic relations, with the exception of antonymy, in this study.

2.2 Wordnet (WN)

The Cognitive Science Laboratory of Princeton University created WN, a lexical knowledgebase for English, in 1990 (Fellbaum, 1998). Synsets (a group of form-meaning pairs sharing same sense) are the main units used in WN to organize the lexicon conceptually. Each sense can be expanded either by gloss or context. It is easy for users to distinguish each sense by simply checking the synonyms, the example sentences or explanation. Nouns, verbs, adjectives and adverbs are the main lexical categories to classify all the lexicons. Such classification of lexicons is based on the principles in psycholinguistics. Besides, the semantic relations of each sense in WN are also expressed like a Word-network. In other words, WN resembles an ontology system and links all the semantic relations of words. Therefore, English WN is not just a lexical knowledgebase but also an ontological system that expresses the semantic relations and the concepts of words.

The current version of WN is Wordnet 2.0, but Wordnet 1.6 is more widely used by the most applications in NLP and linguistic research. Therefore, after considering the compatibility with other applications, we connected the ECTED with Wordnet 1.6. However, we are still working on keeping updating our systems by using the content in the new version of WN. We believe this will keep the information updated and shorten the gap caused by the different versions of WN.

3 Inferring Lexical Semantic Relations for WN and ECTED

As we mentioned above, WN does not only express the knowledge of lexicons but also cover the semantic relations of lexicons. Therefore, in order to present such semantic relations clearly and logically, Huang (2002) proposed to use cross-lingual Lexical Semantic Relations (LSRs) to predict the semantic relations in the target language. The proposed framework is shown in Diagram 1.

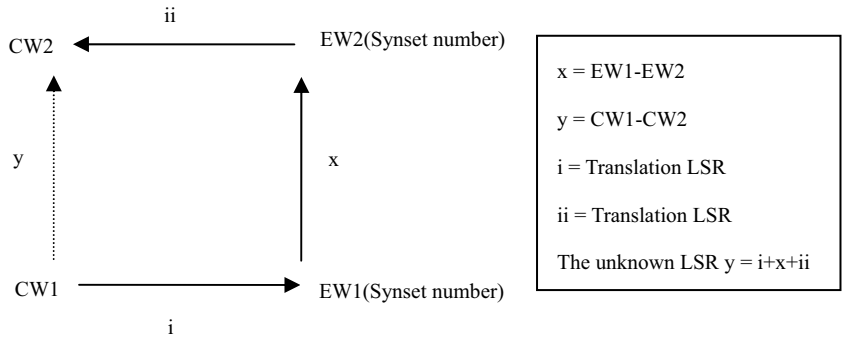

Diagram 1. Translation-mediated LSR (the complete model)

In Diagram1, EW1 and EW2 are head words for two different English synsets. CW1 and CW2 are translation equivalents in ECTED for these two head words. LSR i and ii are the T-LSRs stipulating the semantic relations between the head words and their Chinese TEs. In WN, each synset is linked to a network of their synsets through a number of LSR's. Hence, we use LSR x to represent the semantic relation

between EW1 and EW2. The four LSR's form a closed network that includes three know LSR's: two T-LSRs, i and ii, and one English LSR, x, from WN. The only unknown LSR is y, the semantic relation between CW1 and CW2. Huang et al (2002) claimed that LSR y can be inferred as a functional combination of the three LSRs - i, x and ii.

Language translation does not only involve the semantic correspondences but also the human decision in choosing translation equivalents that are affected by the social and cultural factors. Our main priority in this paper is to infer the lexical semantic information across different language rather than the translational idiosyncrasies, so the elements regarding translational idiosyncrasies are excluded here. In order to simplify the complexity of LSR combination and get a better prediction of LSR, here, we only take account of the situations when LSR ii is exactly equivalent, EW2=CW2 or ii=0. Therefore, we have a reduced model of the translation-mediated LSR Prediction as shown in diagram 2.

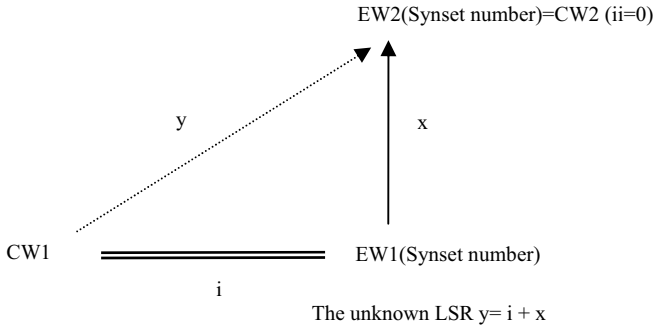

Diagram 2. Translation-mediated LSR (the reduced model)

Synonym, hypernym, hyponym, holonym, meronym and near-synonym are the main semantic relations that we will discuss in the following sections. First of all, we would like to discuss the foundational situation of LSR prediction, synonym, as shown in diagram 3. When translation LSR i is exactly equivalent, i.e. CW1=EW1, and LSR ii is also exactly equivalent, i.e. EW2=CW2, the LSR combination, LSR y, is directly inherited the semantic relation of LSR x.

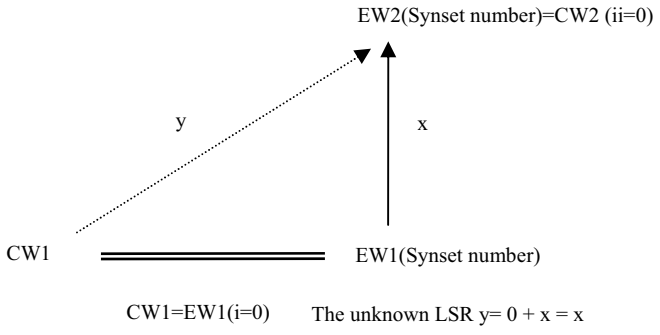

Diagram 3. Translation-mediated LSR (When TEs are synonymous)

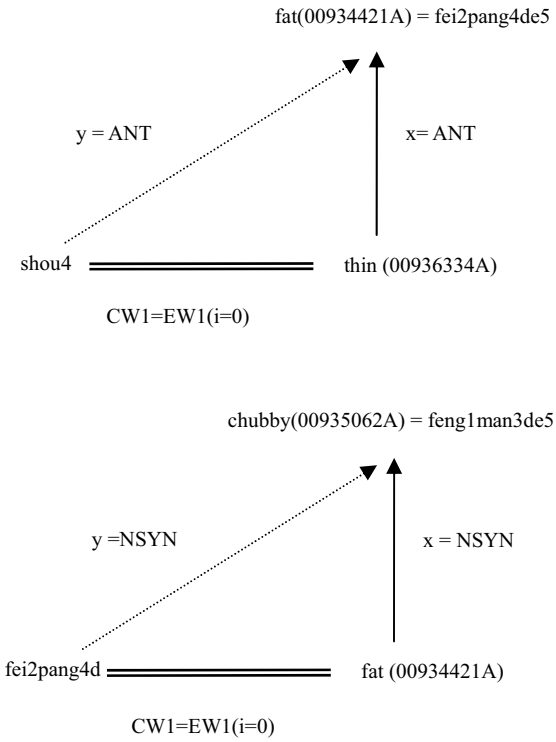

Diagram 4. Examples of the LSR (When TEs are synonymous)

As shown in diagram 4 above, according to the ECTED, the English head word 'thin' is exactly equivalent with 'shou4' in Chinese. The LSR x between EW1 and EW2 in WN is marked 'ANT' which means 'fat' is the antonym of 'thin.' Therefore, according to the prediction in diagram 3, we can infer that the CW2 (fei2pang4de5) is the antonym of CW1 (shou4). The above inference can also be applied to another example in diagram 4. The LSR prediction in WN plays a very crucial role in determining the unknown LSR y. Even an English head word may have more than one sense, it is still very clear to infer the LSR between the TEs. However, there is a potential problem within this inference. If a head word has more than one Chinese TEs which can all correspond to the head word, there might be a problem to consider whether those TEs are really synonyms.

However, the situation is not always that ideal as above. When the Chinese translation equivalents and the corresponded English synset have a non-identical semantic relation, CW1≠EW1, the prediction of LSR y needs to be considered further and carefully.

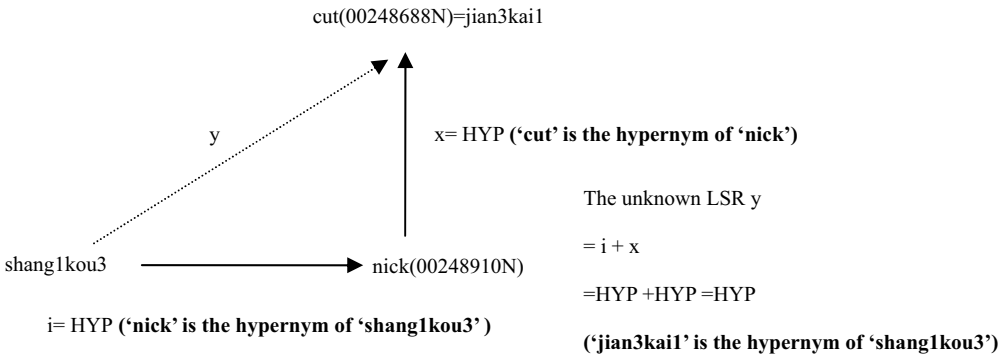

Diagram 5. Predicting LSR (Hypernym) and its example

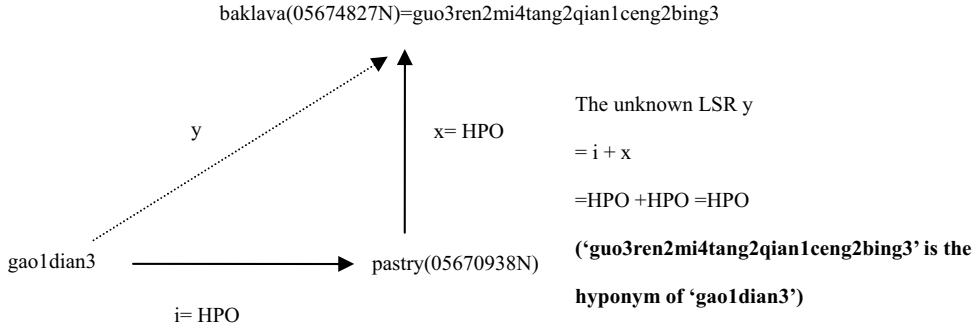

Diagram 6. Predicting LSR(Hyponym) and its example

Logically, hypernym and hyponym are symmetric semantic relations. For instance, if A is a hypernym of B, B is a hyponym of A. For instance, as shown in diagram 5, the English word 'nick' is the hypernym of the Chinese term 'shang1kou3' and 'cut' is the hypernym of 'nick' in WN and the exact translation equivalent of 'cut' in Chinese is 'jian3kai1.' According to the logicality, 'jian3kai1' is the hypernym of 'shang1kou3.' The example of hyponym is shown in diagram 6. Due to the varied semantic relations in WN, the inferences of LSRs , the unknown LSR y = i + x ,for hypernym, hyponym, near-synonym, holonym, and mernoym are listed as below:

Hypernym(HYP)
(a) IF x=ANT
LSR y =HYP +ANT =ANT (CW2 is the antonym of CW1.)
(b) IF x=HYP
LSR y = HYP+HYP =HYP (CW2 is the hypernym of CW1.)
(c) IF x= NSYN
LSR y = HYP+NSYN =HYP (CW2 is the hypernym of CW1.)
(d) IF x = HOL
LSR y = HYP+HOL =HOL (CW2 is the holonym of CW1.)
(e) IF x = all other LSR
LSR y = HYP +all other LSRs = ? (Undecided)

Hyponym(HPO)
(a) IF x=ANT
LSR y =HPO +ANT =ANT (CW2 is the antonym of CW1.)
(b) IF x=HPO
LSR y = HPO+HPO =HPO (CW2 is the hyponym of CW1.)
(c) IF x= NSYN
LSR y = HPO+NSYN =HPO (CW2 is the hyponym of CW1.)
(d) IF x = MER
LSR y = HPO+MER =MER (CW2 is the meronym of CW1.)
(e) IF x = all other LSR
LSR y = HPO +all other LSRs = ? (Undecided)

Near-Synonym(NSYN)
(a) IF x=ANT
LSR y = NSYN+ANT =ANT (CW2 is the antonym of CW1.)
(b) IF x=HYP
LSR y = NSYN+HYP =HYP (CW2 is the hypernym of CW1.)
(c) IF x=HPO
LSR y = NSYN+HPO =HPO (CW2 is the hyponym of CW1.)
(d) IF x= NSYN
LSR y = NSYN+NSYN =NSYN (CW2 is the near-synonym of CW1.)
(e) IF x = MER
LSR y = NSYN+MER =MER (CW2 is the meronym of CW1.)
(f) IF x = HOL
LSR y = NSYN+HOL =HOL (CW2 is the holonym of CW1.)

Holonym(HOL)
(a) IF x=ANT
LSR y = HOL+ANT =ANT (CW2 is the antonym of CW1.)
(b) IF x=HYP
LSR y = HOL+HYP =HYP (CW2 is the hypernym of CW1.)
(c) IF x= NSYN
LSR y = HOL+NSYN =HOL (CW2 is the holonym of CW1.)
(d) IF x = HOL
LSR y = HOL+HOL =HOL (CW2 is the holonym of CW1.)
(e) IF x = all other LSR
LSR y = HPO +all other LSRs = ? (Undecided)

Meronym(MER)
(a) IF x=ANT
LSR y = MER+ANT =ANT (CW2 is the antonym of CW1.)
(b) IF x=HPO
LSR y = MER+HPO =HPO (CW2 is the hyponym of CW1.)
(c) IF x= NSYN
LSR y = MER+NSYN =MER (CW2 is the meronym of CW1.)
(d) IF x = MER
LSR y = MER+MER =MER (CW2 is the meronym of CW1.)
(e) IF x = all other LSR
LSR y = HPO +all other LSRs = ? (Undecided)

4 Implementation and Evaluation

WN 1.6 contains 99,642 English synsets and expands to 157,507 English lemma tokens. On the other hand, the total number of Chinese lemma types found in our ECTED is 108,533. Hence, each Chinese lemma type translates roughly 1.1 English synsets in average.

In comparing the two approaches to parallel wordnet building, we treat at baseline the cases where the translation LSR is synonymy. In others words, these are the cases where both approach I and approach II will make highly accurate predictions (e.g. Huang, et al. 2003). However, if the T-LSR is other than synonymy, we expect the prediction based on source language LSR will be much lower.

In our study, there are in total 372,927 lexical semantic relations that can potentially be bootstrapped when the T-LSR is one of the five semantic relations in study. These are expanded from the following types of translations equivalence relations: 11,396 translation near-synonyms, 2,782 translation hypernyms, 2,106 translation hyponyms, 252 translation meronyms and 145 translations holonyms. For evaluation, due to constraints on resources, we exhaustively check the types with less than 300 lemmas, while randomly checked close to 300 lemmas for the other types.

We first introduce the baseline model where synonym is assumed. This is where source language LSR's will be mapped directly to target languages. We have shown that if the T-LSR is really synonymy, the precision will be 62.7%. However, when the T-LSR's are different, the baseline precision is much lower. In Table 1, such naïve prediction is manually classes into three types: Correct, Incorrect, and Others. 'Correct' means that the prediction is verified. 'Incorrect' means the assigned LSR is wrong. Two scenarios are possible. One is that there is a possible prediction and another one is the correct LSR is different from the predicted one. 'Others' refers to exceptional cases where these is no lexical translation, or the source language LSR is wrongly assigned and so on. Table 1 shows that the baseline for non-synonymous T-LSR is only 47% in average, and range from 30% to 65% for each semantic relation.

	Correct		Incorrect		Others		Total	
NSYN	**400**	**51%**	379	49%	0	0%	779	100%
HYP	**178**	**65%**	72	27%	22	8%	272	100%
HPO	**402**	**40%**	285	28%	330	32%	1017	100%
HOL	**48**	**30%**	108	69%	2	1%	158	100%
MER	**52**	**56%**	32	34%	9	10%	93	100%
Total	**1079**	**47%**	877	37%	363	16%	2319	100%

Table 1 Baseline Results (assuming synonym)

Table 2 shows the comparison between the T-LSR model and the baseline. It shows that there is improvement of 17.8% in average and that there is gain in precision for each LSR type. The improvement varies from just below 2% to 39%.

	Baseline		T-LSR		Difference		Improvement	
NSYN	400	51%	556	71%	156	20 %	156/400	39 %
HYP	178	65%	184	66%	6	2.2%	6/178	3.4%
HPO	402	40%	409	40%	7	0.7%	7/402	1.7%
HOL	48	30%	64	41%	16	10.1%	16/48	33.3%
MER	52	56%	58	62%	6	6.5%	6/52	11.5%
Total	1079	47%	1271	55%	191	8.2%	191/1080	17.7%

Table 2 Precision of using the LSR inferences

5 Conclusion

It is interesting to note that the classes with least improvements are hypernymy and hyponymy. Since these are the classical IS-A relations, we hypothesize that their predictions are similar to the baseline relation of synonym. If we take these two relations out, the T-LSR model with inference rules has a precision difference of 17.3% (178/1030), as well as an improvement of 35.6% (178/500). These are substantial improvements over the baseline model. The result will be reinforced when the evaluation is completed. We will also analyze the prediction based on each T-LSR to give a more explanatory account as well a measure confidence or prediction. The result offers strong support for T-LSR as a model for bootstrapping parallel wordnets with a low-density target language.

References

Fellbaum, C. (ed.) 1998. *Wordent: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Huang, Chu-Ren, D.B. Tsai, J.Lin, S. Tseng, K.J. Chen and Y. Chuang. 2001 *Definition and Test for Lexical Semantic Relation in Chinese.* [in Chinese] Paper presented at the Second Chinese Lexical Semantics Workshop. May 2001, Beijing, China.

Huang, Chu-Ren, I-Ju E. Tseng, Dylan B.S. Tsai. 2002. *Translating Lexical Semantic Relations: The first step towards Multilingual Wordnets.* Proceedings of the COLONG2002 Workshop "SemaNet:Building and Using Semantic Networks", ed. By Grace Ngai, Pascale Fung, and Kenneth W. Church, 2-8.

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*. pp.509-531.

Pianta, Emanuel, L. Benitivogli, C. Girardi. 2002 *MultiWordnet: Developing an aligned nultilingual database.* Proceedings of the 1st International WordNet Conference, Maysore, Inda, pp.293-302.

Tsai, D.B.S., Chu-Ren Huang, J.Lin, K.J. Chen and Y. Chuang. 2002. *Definition and Test for Lexical Semantic Relation in Chinese.* [中文詞義關係的定義與判定原則] *Journal of Chinese Information Processing* [中文信息學報]. 16.4.21-31.

Vossen P. (ed.). 1998. *EuroWordNet: A multilingual database with lexical semantic networks.* Norwell, MA: Kluwer Academic Publisher