Chinese WordNet Domains: Bootstrapping Chinese WordNet with Semantic Domain Labels *

Lung-Hao Lee[a], Yu-Ting Yu[a], and Chu-Ren Huang[a,b]

[a]Institute of Linguistics, Academia Sinica,
128 Academia Road, Section 2, Taipei 115, Taiwan
{lunghao, tina0822, churen}@gate.sinica.edu.tw
[b]Department of Chinese & Bilingual Studies, The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong
churen.huang@inet.polyu.edu.hk

Abstract. We bootstrapped Chinese WordNet with semantic domain labels of WordNet Domains for constructing a language resource called Chinese WordNet Domains. The bootstrapping methods work from three aspects: 1) Princeton WordNet alignment, 2) lexical semantic relations and 3) domain taxonomy mapping. Experimental results of our proposed bootstrapping based domain predication achieve satisfying effects. We believe the resulting Chinese WordNet Domains will be the first oriental language resource, which can be used to interoperate with the existing WordNet Domains of several languages and benefits for cross-language and domain-specific researches and applications. In addition, we also plan to release resulting Chinese WordNet Domains to the community for research purposes.

Keywords: Bootstrapping, Chinese WordNet, WordNet Domains, Multi-label.

1 Introduction

Princeton WordNet is an English lexical database that groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms, which are named as synsets (Fellbaum, 1998; Miller, 1995). The Global WordNet Association (GWA), built on the results of Princeton WordNet and Euro WordNet (Vossen, 2004), is a free and public association that provides a platform that shares and connects all languages in the world. For Mandarin Chinese in Taiwan, Huang et al. (2004a) constructed the Academia Sinica Bilingual Ontological Wordnet (Sinica BOW), which integrates WordNet, English-Chinese Translation Equivalents Database (ECTED) and SUMO for cross-language linguistic studies. As a follow-up, Chinese WordNet has been built as a robust lexical knowledge system which embodies a precise expression of sense and sense relations as well (Huang et al., 2008b). In recent years, WordNet-like resources have become one of the most reliable and essential resource for linguistic studies for all languages (Niles and Pease, 2003; Budanitsky and Hirst, 2006; Soria et al., 2009a).

Semantic domain labels, characterized by domain-specific lexica, are profitably used to describe texts and word senses according to general subjects, such as sport, finance, and politics. WordNet Domains (Magnini and Cavaglia, 2000) was created by extending the Princeton WordNet with domains labels. Synsets have been semi-automatically annotated with at least one domain label. A domain can include synsets of different part-of-speech and from different WordNet sub-hierarchies. So far the existing WordNets such as Italian WordNet,

* This work was funded by National Science Council, Taiwan under Grants NSC 97-2923-I-001-001-MY3, and also cooperated with EU-FP7 KYOTO project.

Spain WordNet, Hebrew WordNet, and Romanian WordNet are annotated with the same domains labels from WordNet Domains. WordNet Domains has been viewed as an important language resource for domain-based language processing (Magnini et al, 2002a; 2002b; Gliozzo, 2006).

In this study, we use bootstrapping methods to automatically annotate word senses of Chinese WordNet using semantic labels of WordNet Domains. The bootstrapping methods work from three aspects: 1) Princeton WordNet alignment, 2) lexical semantic relations and 3) domain taxonomy mapping. We believe the resulting Chinese WordNet Domains will be the first oriental language resource, which can be used to interoperate with existing of WordNet Domains of several languages and benefit for cross-language and domain-specific researches and applications.

The rest of this paper is organized as follows: Section 2 introduces related studies on Chinese WordNet and WordNet Domains. Section 3 describes the bootstrapping methods for automatically annotating semantic domain labels using Chinese WordNet and WordNet Domains. Performance evaluation and experimental results are presented in Section 4. We discuss the resulting Chinese WordNet Domains and its application scenarios in Section 5. Section 6 concludes this study with future research

2 Related Work

The section is devoted to relevant studies on Chinese WordNet and WordNet Domains.

2.1 Chinese WordNet

Creating a semantic relation-based language resource is a time consuming and labor intensive task, especially for Chinese, due to the unobvious definition and distinction among characters, morphemes and words. Chinese WordNet[1] (CWN) is built by Academia Sinica and has successively extended its scope so far. Lemmas included in CWN mainly fall on the medium frequency words. Each lexical entry is analyzed according to the guidelines of Chinese word sense distinctions (CKIP, 2003; Huang et al. 2003a) which contain information including Part-of-Speech, sense definition, example sentences, corresponding English synset(s) from Princeton WordNet, lexical semantic relations and so on. Unlike Princeton WordNet, CWN has not been constructed mainly on the synsets and semantic relations. Rather, it focuses to provide precise expression for the Chinese sense division and the semantic relations needs to be based on the linguistic theories, especially lexical semantics (Huang et al., 2008b). Moreover, Huang et al. (2005a) designed and implemented the Sinica Sense Management System (SSMS) to store and manage word sense data generated in the analysis stage. SSMS is meaning-driven. Each sense of a lemma is identified specifically using a unique identifier and given a separate entry. There are 8,628 lemmas /25,938 senses analyzed and stored in SSMS until December 2008. Lee et al (2009) used WordNet-LMF (Soria et al., 2009b) to represent lexical semantics in Chinese WordNet. The compiled CWN-LMF will be released to the community for linguistic researches. Figure 1 shows the result of the noun 植物 (zhi2 wu4, plant) in Chinese WordNet.

Huang et al. (2004b) proposed Domain Lexico-Taxonomy (DLT) as a domain taxonomy populated with lexical entries. By using DLT with Chinese WordNet and Domain Taxonomy, there were 2,541 Chinese senses that are linked with and distributed in 141 domain nodes. In addition, Huang et al. (2005b) further applied DLT approach to a Chinese thesaurus called CiLin and showed with evaluation that DLT approach is robust since the size and number of domain lexica increased effectively.

[1] Chinese WordNet, available online at http://cwn.ling.sinica.edu.tw/

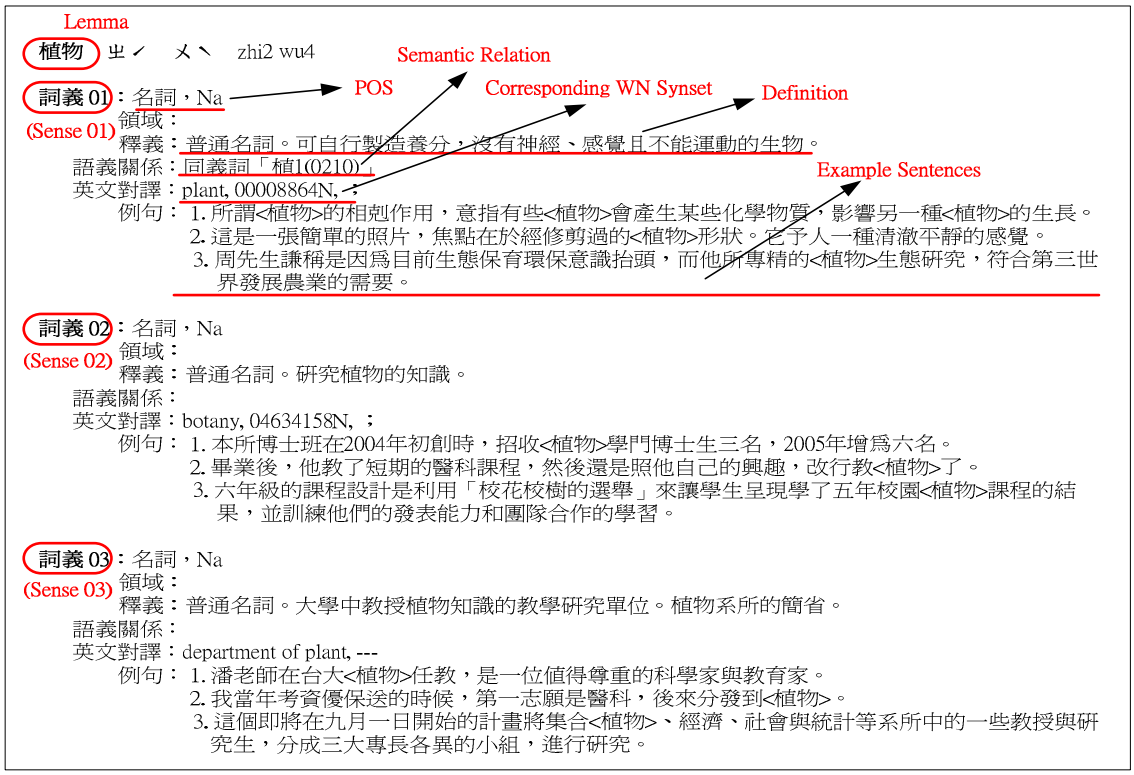

Figure 1: The three senses of the noun 植物 (zhi2 wu4, plant) return by Chinese WordNet.

2.2 WordNet Domains

WordNet Domains[2] is a linguistic resource constructed by ITC-IRST where the Princeton WordNet is augmented with domain labels (Magnini and Cavaglia, 2000). Synsets have been semi-automatically annotated with at least one domain labels. These domain labels, such as Music, Transport, and Law, are selected from a set of about 200 labels that are hierarchally organized referred to the Dewey Decimal Classification (DDC). Magnini and Cavaglia (2000) manually annotated a small number of high level synsets with their domain labels. Then, an automatic procedure exploited some of the WordNet relations to extend the manual assignment to the reachable synsets. In addition, an exception procedure was used to prevent a wrong propagation. Bentivogli et al. (2004) further revised the WordNet Domains Hierarchy (WDH) with a clear semantics and evaluated the coverage and balancing of Basic Domains of WDH. The latest version, WordNet Domains 3.2, contains the mapping between Princeton WordNet 2.0 synsets and their corresponding domains. 45 Basic Domains of total 168 domains are used to annotate WordNet synsets. Take “00197005-n history law” for example, “00197005-n” is the synset off set and Part-of-Speech and “history law” is the list of domains associated to the synset. Notice that an additional label named as “Factotum” was assign to Generic synset, which was hard to classify in a specific domain and Stop senses synsets, which appeared frequently in different contexts, such as colors, numbers, week days and so forth. So far the existing Wordnets such as Italian Wordnet, Spain WordNet, Hebrew Wordnet, and Romanian Wordnet are annotated with the same domain labels from WordNet Domains. Moreover, WordNet Domains has been exploited in the framework of MultiWordNet (Pianta et al., 2002) and considered a crucial language resource for NLP tasks, such as Word Sense Disambiguation (Gliozzo et al. 2004; Kolte and Bhirud, 2008; Magnini et al. 2002b) and Text Categorization (Katsiouli et al. 2007; Vázquez et al. 2006).

[2] WordNet Domains, available online at http://wndomains.itc.it/wordnetdomains.html

3 Bootstrapping Semantic Domain Labels

The section describes our proposed bootstrapping methods using Chinese WordNet and WordNet Domains. Chinese WordNet focuses to provide precise expression for the Chinese sense division and lexical semantic relations. In addition, partial senses are annotated with domain nodes from Domain Lexico-Taxonomy. WordNet Domains contains the mapping between Princeton WordNet 2.0 synsets and their corresponding domains. Synsets have been semi-automatically annotated with at least one domain label, which is selected from Dewey Decimal Classification (DDC).

Since cross-lingual lexical semantic relation inferences were examined by bootstrapping a Sinica BOW with Princeton WordNet (Huang et al. 2002; 2003b), we decide to use bootstrapping methods for constructing a language resources named as Chinese WordNet Domains. By using an existing WordNet Domains as a medium, we automatically annotate word senses of Chinese WordNet with semantic domain labels from three aspects: 1) Princeton WordNet alignment, 2) lexical semantic relations and 3) domain taxonomy mapping. Details will be described as the following subsections.

3.1 Alignment-mediated Domain Prediction

Word senses of Chinese WordNet are strictly aligned with the corresponding synsets of Princeton WordNet; therefore, we can use alignment-mediated information to bootstrap a Chinese version of WordNet Domains. Figure 2 demonstrates the bootstrapping method based on alignment-mediated domain prediction: CW_1 stands for the Chinese sense which can be aligned to English sense, EW_1, through equal synonymy relation. DDC represents the Dewey Decimal Classification where the domain labels are selected from. If EW1 is annotated with at least one semantic domain selected from DDC, and EW_1 is aligned to CW_1, then CW_1 can be predicated with the same domain labels of EW_1. For example, "04071401-n" is the first sense of the lemma 愛迪生 (Edison), which is aligned to Princeton WordNet 2.0 synset "10235982-n". In WordNet Domains, "person" is the list of domains associated to the synset "10235982-n". Base on alignment-mediated domain prediction, 04071401-n ("愛迪生_1") of Chinese WordNet will be annotated with domain "person".

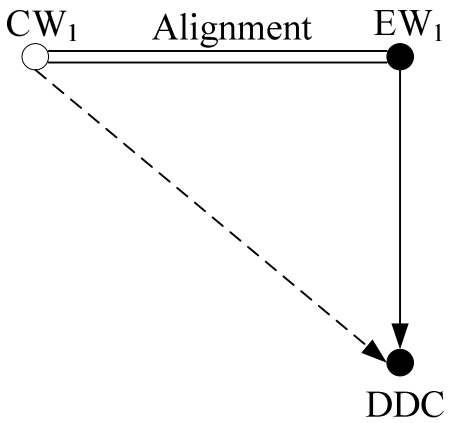

Figure 2: Alignment-mediated domain prediction.

3.2 LSR-mediated Domain Prediction

In Chinese WordNet, several lexical semantic relations, including synonym, hypernym, hyponym, antonym and so forth, are manually annotated to the corresponding senses. If a sense has been annotated with domain labels of WordNet Domains based on alignment-mediated

domain prediction, we can further infer the domain labels to senses which are not tagged through well-defined lexical semantic relations.

Figure 3 illustrates Lexical Semantic Relations (LSRs) based domain prediction. CW_1 represents the Chinese sense that is annotated with semantic domains from DDC by alignment-mediated domain prediction. If there are existing LSRs between CW_1 and CW_2, CW_2 will be predicated with the same domain labels of CW_1 through LSRs. We used four LSRs of Chinese WordNet to infer domain labels, i.e. synonymy, near- synonymy, paranymy and variants. Among these LSRs, the semantic relation paranymy is used to refer to relation between any two lexical items belonging to the same semantic classification (Huang et al. 2008a). For example, the set of "spring/summer/fall/winter" has paranymy relation of main concept of "seasons in a year". Furthermore, the relation variants denote the variation of Chinese characters (Hong et al. 2005). For instance, "爲什麼" and "爲甚麼" both represent the meaning "why" but use different writing notation to stand for the second character. Take the second sense of the lemma "水瓶座" (Aquarius) for example, since "05085502-n" (水瓶座_2) has been annotated as "astrology" from alignment-mediated domain prediction, and "05181002-n" (摩羯座_2, Capricorns) is the paranymy of "05085502-n", so the domain labels of "05181002-n" is determined as "astrology" based on LSR-mediated domain prediction.

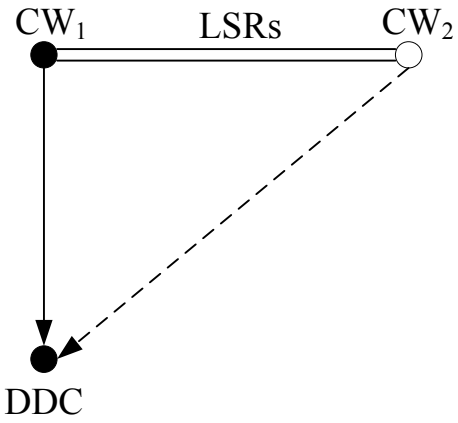

Figure 3: LSR-mediated domain prediction.

3.3 Mapping-mediated Domain Prediction

So far some senses of Chinese WordNet are annotated with domain nodes from Domain Lexico-Taxonomy (DLT) (Huang et al. 2004b). If the mapping of domain nodes of DLT and semantic domains from DDC is well-defined, we can annotate the senses with semantic domains of WordNet Domains based on mapping-mediated domain prediction.

Figure 4 demonstrates the bootstrapping method based on mapping-mediated domain prediction: CW_1 represents the Chinese sense which is annotated with domain nodes from DLT. If the mapping of DLT and DDC can be constructed, CW_1 will be predicated with domain labels of DDC through DLT-mapping. For example, "06736101-n" is the first sense of the first lemma 佛 (Buddha), and "06736101-n" (佛 1_1) is annotated with the domain node 佛教 (Buddhism). After mapping of DLT and DDC, 佛教 (Buddhism) is mapped into "religion", so "06736101-n" is annotated with "religion" from WordNet Domains.

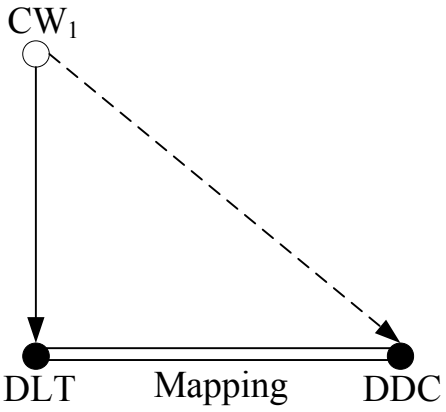

Figure 4: Mapping-mediated domain prediction.

4 Experiments and Performance Evaluation

4.1 Data Sets

We used Chinese WordNet 1.6 and WordNet Domains 3.2 to bootstrap the resulting Chinese WordNet Domains. In Chinese WordNet 1.6, 8,628 lemmas / 25,938 senses have been analyzed. Among these senses, there are 18,789 synonyms, 1,801 near-synonyms, 3,029 paranyms and 923 variants. In addition, 2,541 senses in Chinese WordNet 1.6 have been annotated with domain labels from DLT. WordNet Domains 3.2 contains the mapping between Princeton WordNet 2.0 synsets and their corresponding domains. 45 Basic Domains of total 168 domains are used to annotate WordNet synsets. 115,424 synsets in Princeton WordNet 2.0 synsets have been annotated with at least one domain labels from DDC. Among these annotated synsets, 40,995 synsets have domain label, "Factotum," representing synsets that do not belong to any specific semantic domains.

4.2 Evaluation Criteria

For each sense in Chinese WordNet 1.6, the bootstrapping methods generate a list of labels – the semantic domains to which the sense possibly belongs. In order to measure the performances of our proposed methods, we used the next three measures: multi-label precision, multi-label recall, and multi-label F-measure, which are well-known in multi-label classification problem. We defined these measures as the following equations with notations adapted to our problem.

$$\Pr ecision = \frac{\left\| correct\_labels \cap predicated\_labels \right\|}{\left\| predicated\_labels \right\|} \qquad (1)$$

$$\mathrm{Re} call = \frac{\left\| correct\_labels \cap predicated\_labels \right\|}{\left\| correct\_labels \right\|} \qquad (2)$$

$$F - measure = \frac{2 * \Pr ecision * \mathrm{Re} call}{\Pr ecision + \mathrm{Re} call} \qquad (3)$$

Where $\|predicated\_labels\|$ denotes the number of a list of labels predicated for a sense by our proposed bootstrapping methods; $\|correct\_labels\|$ is the number of known correct list of labels for a sense; $\|correct\_labels \cap predicated\_labels\|$ is the number of correctly predicated list of labels for a sense. For example, if well known correct list of labels of a sense is A, B, C, D and our proposed method predicated as B, C, E, in this case, $\|predicated\_labels\|$ is 3 (i.e. B, C, E); $\|correct\_labels\|$ is 4 (i.e. A, B, C, D); $\|correct\_labels \cap predicated\_labels\|$ is 2 (i.e. B, C), so the precision is 0.67 and the recall is 0.5. A precision of 0.67 means that at least two labels is correct if our proposed method predicated a list of 3 labels as output. Similarly, we get a recall of 0.5 when a half of domain labels can be correctly predicted. We use macro-averaging precision, recall, and F-measure as our experimental evaluation criteria.

4.3 Experimental Results

We manually annotated the senses which are not predicated as "Factotum" as ground truth for performance evaluation. Table 1 shows the performance evaluation of our proposed bootstrapping methods. Our proposed mapping-mediated domain predication achieved the best results, which scored a precision of 97.15%, a recall of 96.33% and an F measure of 96.6%. In addition, experimental results indicated alignment-mediated domain predication scored a precision of 83.86%, a recall of 82.72%, and an F measure of 82.46% as compared to a precision of 70.4%, a recall of 69.83%, and a F measure of 69.54% for LSR-mediated domain predication.

Table 1: Performance evaluation of our proposed bootstrapping methods.

Methods	#Sense	Precision (%)	Recall (%)	F measure (%)
Alignment-mediated	6,669	83.86	82.72	82.46
LSR-mediated	174	70.4	69.83	69.54
Mapping-mediated	2,350	97.15	96.33	96.6

5 Discussion

Experimental results indicated mapping-mediated domain prediction can achieve the best effects. It is because domain taxonomy mapping is more clarifying if the senses are already manually annotating with original domain taxonomy. On the other hand, LSR-mediated domain prediction achieved the worse effect. The main reason for this case is that LSR-mediated is based on alignment-mediated domain prediction method. Therefore, it cannot obtain better results than alignment-mediated bootstrapping. Since manually constructed language resources contain inconsistencies and errors, our proposed methods use Chinese WordNet and WordNet Domains to automatically bootstrap semantic domain labels obtain at least satisfying effects.

We can further use the resulting Chinese WordNet Domains as the tool for quality assurance of Chinese WordNet and WordNet Domains. For example, we can check the inconsistencies of writing variants of the same sense with the different semantic domain labels to verify the correctness of lexical semantic relations and sense alignment to Princeton WordNet in Chinese WordNet. Moreover, we can interoperate and exchange linguistic information in Chinese WordNet Domains with WordNet Domains of several languages to form cross-language resources for research purposes. In addition, we can put the senses with the same domain label together as domain-specific lexicons for domain-specific applications.

6 Conclusions

We bootstrapped Chinese WordNet with semantic domain labels using WordNet Domains to construct a language resource called as Chinese WordNet Domains. The bootstrapping methods work from three aspects: 1) Princeton WordNet alignment, 2) lexical semantic relations and 3) domain taxonomy mapping. Experimental results indicated our proposed bootstrapping based domain predication achieve satisfying effects. We believe the resulting Chinese WordNet Domains will be the first oriental language resource, which can be used to interoperate with existing of WordNet Domains of several languages and benefit for cross-language and domain-specific researches and applications.

Future work is investigated along several directions. An attempt to use Chinese WordNet Domains as external resources for domain-specific semantic search is ongoing. In addition, since WordNet of several languages have been annotated with the same domain labels, cross-language information retrieval using Chinese WordNet Domains will also be investigated. Finally, we also plan to release resulting Chinese WordNet Domains to the community for research purposes.

References

Bentivogli, L., P. Forner, B. Magnini and E. Pianta. 2004. Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. *Proceedings of COLING Workshop on Multilingual Linguistic Resources*.

Budanitsky, A. and G. Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1), 13-47.

CKIP. 2003. *Sense and Sensibility Vol. I*. Technical Report 03-01. Taipei: Academia Sinica.

Fellbaum, C. 1998. *WordNet: an Electronic Lexical Database*. The MIT Press.

Gliozzo, A., C. Strapparava and I. Dagan. 2004. Unsupervised and Supervised Exploitation of Semantic Domains in Lexical Disambiguation. *Computer Speech and Language*, 18(3), 275-299.

Gliozzo, A. 2006. Semantic Domains and Linguistic Theory. *Proceedings of the LREC Workshop on Toward Computational Models of Literary Analysis*.

Hong, J.-F., C.-R. Huang and Y.-J. Wu. 2005. Towards a Study on the lexical Semantics of Character- and Word- Variants. *Proceedings of the 6th Chinese Lexical Semantics Workshop*.

Huang, C.-R., B. S. Tsai, C.-X. Weng, N.-X. Chu, W.-R. Ho, L.-W. Huang and I.-N. Tsai. 2003a. Sense and Meaning Facet: Criteria and Operational Guidelines for Chinese Sense Distinction. *Proceedings of the 4th Chinese Lexical Semantics Workshop.*

Huang, C.-R., C.-L. Chen, C.-X. Weng, X.-B. Li, Y.-X. Chen and K.-J. Chen. 2005a. The Sinica Sense Management System: Design and Implementation. *Computational Linguistics and Chinese Language Processing*, 10(4), 417-430.

Huang, C.-R., I.-J. Tseng and B. S. Tsai. 2002. Translating Lexical Semantic Relations: The First Step towards Multilingual WordNets. *Proceedings of the COLING Workshop on SemaNet: Building and Using Semantic Networks.*

Huang, C.-R., I.-J. Tseng, B. S. Tsai and B. Murphy. 2003b. Cross-lingual Portability of Semantic Relations: Boostrapping Chinese WordNet with English WordNet Relations. *Language and Linguistics*, 4(3), 509-532.

Huang, C.-R., I.-L. Su, P.-Y. Hsiao and X.-L. Ke. 2008. Paranymy: Enriching Ontological Knowledge in WordNets. *Proceedings of the 4th Global WordNet Conference*.

Huang, C.-R., R.-Y. Chang and X.-B. Li. 2004a. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. *Proceedings of the 4th International Conference on Language Resources and Evaluation*.

Huang, C.-R., S.-K. Hsieh, J.-F. Hong, Y.-Z. Chen, I-L. Su, Y.-X. Chen and S.-W. Huang. 2008b. Chinese Wordnet: Design, Implementation, and Application of an Infrastructure for Cross-lingual Knowledge Processing. *Proceedings of the 9th Chinese Lexical Semantics Workshop*.

Huang, C.-R., X.-B. Li and J.-F. Hong. 2004b. Domain Lexico-Taxonomy: an Approach Towards Multi-domain Language Processing. *Proceedings of the Asian Symposium on Natural Language Processing to Overcome Language Barriers*.

Huang, C.-R., X.-B. Li and J.-F. Hong. 2005b. The Robustness of Domain Lexico-Taxonomy: Expanding Domain Lexicon with Cilin. *Proceedings of the 4th ACL SIGHAN Workshop on Chinese Language Processing.*

Katsiouli, P., V. Tsetsos and S. Hadjiefthymiades. 2007. Semantic Video Classification Based on Subtitles and Domain Terminologies. *Proceedings of SAMT Workshop on Knowledge Acquisition from Multimedia Content*.

Kolte, S. G. and S. G. Bhirud. 2008. Word Sense Disambiguation Using WordNet Domains. *Proceedings of the 1st IEEE International Conference on Emerging Trends in Engineering and Technology*.

Lee, L.-H., S.-K. Hsieh and C.-R. Huang. 2009. CWN-LMF: Chinese WordNet in the Lexical Markup Framework. *Proceedings of the 7th Workshop on Asian Language Resources.*

Magnini, B. and G. Cavaglia. 2000. Integrating Subject Field Codes into WordNet. *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Magnini, B., C. Strapparava, G. Pezzulo and A. Gliozzo. 2002a. Comparing Ontology-Based and Corpus-based Domain Annotations in WordNet. *Proceedings of the 1st International WordNet Conference*.

Magnini, B., C. Strapparava, G. Pezzulo and A. Gliozzo. 2002b. The Role of Domain Information in Word Sense Disambiguation. *Natural Language Engineering*, 8(4), 359-373.

Miller, G. A. 1995. WordNet: a Lexical Database for English. *Communications of the ACM*, 38(11), 39-41.

Niles, I. and A. Pease. 2003. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. *Proceedings of the International Conference on Information and Knowledge Engineering*.

Pianta, E., L. Bentivogli and C. Girardi. 2002. MultiWordNet: Developing an Aligned Multilingual Database. *Proceedings of the 1st Global WordNet Conference*.

Soria, C., M. Monachini, F. Bertagna, N. Calzolari, C.-R. Huang, S.-K. Hsieh, A. Marchetti and M. Tesconi. 2009a. Exploring Interoperability of Language Resources: the Case of Cross-lingual Semiautomatic Enrichment of Wordnets. *Language Resources and Evaluation*, 43(1), 87-96.

Soria, C., M. Monachini and P. Vossen. 2009b. Wordnet-LMF: Fleshing out a Standardized Format for Wordnet Interoperability. *Proceedings of the ACM Workshop on Intercultural Collaboration*.

Vázquez, S., Z. Kozareva and A. Montoyo. 2006. Textual Entailment beyond Semantic Similarity Information. *Lecture Notes in Computer Science*, 4293, 900-910.

Vossen, P. 2004. EuroWordNet: a Multilingual Database of Autonomous and Language-specific Wordnets Connected via an Inter-Lingual-Index. *International Journal of Linguistics*, 17(2), 1-23