Induction of Classification from Lexicon Expansion : Assigning Domain Tags to WordNet Entries

Echa Chang*, Chu-Ren Huang, Sue-Jin Ker***, Chang-Hua Yang*****
*University of Waterloo, 200 University Ave. W., Waterloo, ON N2L 3G1 Canada
cecha@yahoo.com
****Institute of Linguistics, Academia Sinica, Nankang, Taipei, 115, Taiwan
churen@sinica.edu.tw
*****Soochow University
ksj@sun.cis.scu.edu.tw, changhua@mail2000.com.tw

Abstract

We present in this paper a series of induced methods to assign domain tags to WordNet entries. Our prime objective is to enrich the contextual information in WordNet specific to each synset entry. By using the available lexical sources such as Far East Dictionary and the contextual information in WordNet itself, we can find a foundation upon which we can base our categorization. Next we further examine the similarity between common lexical taxonomy and the semantic hierarchy of WordNet. Based on this observation and the knowledge of other semantic relations we enlarge the coverage of our findings in a systematic way. Evaluation of the results shows that we achieved reasonable and satisfactory accuracy. We propose this as the first step of wordnet expansion into a bona fide semantic network linked to real-world knowledge.

0. Introduction[1]

WordNet is a lexicon comprising of nouns, verbs, adjectives and adverbs. Its basic organization is based on different semantic relations among the words. Entries (or lemas) sharing the same meaning is grouped into a synset and assigned with a unique sense identification number for easy retrieval and tracking purposes. This unique offset number gives the information about the parts of speech and the hierarchy position to which a specific synset belongs. For nouns and verbs the synsets are grouped into multiple lexical hierarchies; modifiers such as adjectives and adverbs are simply "organized into clusters on the basis of binary opposition (antinomy)." [1] This lexical hierarchy makes the lexical domain assigning task more straightforward because it coincides with a ontological taxonomy in many aspects. The primary objective of our project is to enrich the WordNet knowledge content due to the fact that "WordNet lacks relations between related concepts." [2] We adopt WordNet itself, together with other lexical resources to develop an integrated domain specific lexical resource.

1. The Five Tagging Methods

Starting with two lexical resources, we employed five steps to assign and expand domain tags. Basically, the explicit domain information from Far East Dictionary as well as WordNet's own hierarchy of semantic relation are used to extend the coverage of domain - assignment.

1.1 Domain Data Lookup from Far East Dictionary

The digital file of Far East Dictionary contains complete information for each word entry that can be found in an ordinary printed version. Most of all, it lists the domain

[1] This research is partially funded by an IDLP project grant from the National Science Council of Taiwan, ROC. Work reported in this paper was carried out in summer 2001, during Chang's internship at Academia Sinica. We are indebted to two anonymous reviewers of SemaNet 2002, as well as from the First International WordNet Conference for their helpful comments. An earlier version of this paper was accepted by the first IWC but was not presented because of the authors' travelling difficulties at that time. We thank colleagues at Academia Sinica, especially Shu-Chuan Tseng, Keh-jiann Chen, and members of the WordNet group, for their input and help.

information for each vocabulary wherever possible. Thus we employ the available data from a text source file (each vocabulary entry is organized as one single row) and extract all the information by running a string manipulation program coded in Visual Basic. During the extraction process we only take into account the part of speech of each word in Far East Dictionary. Next, we map the domains obtained from Far East Dictionary if the word and its part of speech coincides with the entries in our database which contains a complete list of synset. Since WordNet collects only nouns, verbs, adjectives and adverbs, we only extract the domain data that falls into these four categories. Later we group the information in a database table and extent the assigned domains of each word to its synset. Table 1 is an example of our database table which 'contains all the adverbial uses of `aback.'

id	term	domain
00073303R	aback	aviation,
00073386R	aback	aviation,

Table 1 Example of The Far East Dictionary Domain Database Table

In Table 1, it is shown that 'aback' has two adverbial senses. Since in Far East Dictionary "aback" is labeled with domain 'aviation,' extra work of expansion is necessary to further label all of its adverb synset with the same domain to maintain the integrity of the information. Because both the extraction and expansion method would produce ambiguities in domain assignment, manual verifications are required in the future.

1.2 Extracting Domain Information from WordNet Sense Description

Each WordNet entry (i.e. each synset) is followed by its sense. Although there is no specifically defined set of controlled vocabulary, the sense definition does specify the field the synset members are commonly used in that specific field of study, such as biology, physics or chemistry. This specification comes in a special format contained in a bracket for each WordNet entry so that extraction of data is possible and straightforward. Due to the fact that each domain is directly extracted by its corresponding synset, there is simply no ambiguity in assigning the domain tags. And if there is more than one lexical item in that synset, all will share the same domain tag.

1.3 Establishment of a Common Domain Taxonomy for Nouns

Each lexical resource uses a different domain taxonomy, which may be explicitly defined or implicitly assumed. Hence, when combining domain information from multiple sources, the establishment of a Common Domain Taxonomy (CDT) is crucial for both efficient representation as well as effective knowledge merging. Our survey of existing domain taxonomy, including LDOCE, HowNet, Tongyici Cilin, etc., show that there is quite a lot in common. Hence we decide to build a working CDT based on the two resources we have. Note that since our goal is to establish a domain taxonomy for wordnets (for English now and for Chinese in the future), the existing domain information in WordNet need to be assumed as defaults that can be over-ridden. Hence a model of CDT based on basic binary combination involving WordNet is necessary.

After collecting all the domain tags from the two resources, we build our CDT. First, all common domain nodes are put in a hierarchy based on their relation. Second, inconsistent domain names are resolved. Last, when gaps appear after all domain tags are attached to the taxonomy, new domain categories are adopted to fill in the gaps and make a more complete CDT. Since top taxonomy presupposes a particular view on conceptual primacy and may differ in different lexical sources, we took a bottom-up approach to our CDT. That is, right now each taxonomy tree now stops at some broad-consensus level without being committed to a higher taxonomy. The following is a partial list of our current CDT.

Humanity
【Linguistics】
【Rhetorical Device】
【Literature】
【History】
【Archeology】
…
Social Science
【Sociology】
【Statistics】
【Economics】
【Business】
【Finance】
…
Formal Science
【Mathematics】

【Geometry】
【Algebra】
…

Natural Science

【Physics】
【Nuclear】
【Chemistry】
【Biology】
【Palaeontology】
【Botany】
【Animal】
【Fish】
【Bird】
…

Applied Science

【Medicine】
【Anatomy】
【Physiology】
【Genetics】
【Pharmacy】
【Agriculture】
…

Fine Arts

【Painting】
【Sculpture】
【Architecture】
【Music】
【Drama】
…

Entertainment

【Sports】
【Balls】
【Track & Field】
【Competition】
【Game】
【Board】
【Card】
…

Proper Noun

【Name】
【Geographical Name】
【Country】
【Religion】
【Trademark】
…

Humanity

【Archaic】
【Informal】
【Slang】
【Metaphor】
【Formal】
【Abbreviate】
…

Lexical Sources

【Latin】
【Greece】
【Spanish】
【French】
【American】
…

Please note that by induction and actual examples from the lexical organization in WordNet, it is found that a hyponym is very likely to belong to the same domain as its hypernym. Similar results are also found for wordnet based cross-lingual inference of lexical semantic relations [4]. For instance, under the term 'mathematics,' all the hyponyms below are related to this field of study. To make us of this lexical semantic phenomenon, we make a table of all the domain terms and map them to their unique WordNet sense identification number. Later we use the tree expansion method (discussed in more detail in Section 2.4) to trace down all the hyponyms. For example, by using this method, the hyponyms of 【Linguistics】 are all labeled as 'linguistics' and so forth.

1.4 Lexical Hierarchy Expansion of Nominal Domain Assignment

WordNet has is a lexical semantic hierarchy linking all synsets with lexical semantic relations. We convert all the relations to a database in a relational table, as shown in Table 2 [1]:

Hypernym ID	Hyponym ID	Relation
00001740A	04349777N	=
00001740A	00002062A	!
00001740N	00002086N	~
…	…	…

Table 2 Lexical Relation Table

The relation symbols in Table 2 are adopted from the WordNet database files. These symbols are saved with each synset entry to indicate a specific semantic relation with other synsets. The implemented information allows us to trace and locate all the related synsets.

WN Relation Symbol
Antonym: !
Hyponym: ~
Hypernym: @
Meronym: #
Holonym: %
Attribute: =

Table 3 Relations and Pointer Symbols

By manipulating Table 2 with SQL, all nouns can be traced to the eight unique beginners.

Unique Beginners of Nouns In WordNet
Entity,something
Abstraction
Act,human action,human activity
State
Event
Group,grouping
Phenomenon
Possession

Table 4. The Eight Unique Beginners for Nouns

The general structure of tree expansion can be visualized as Figure1:

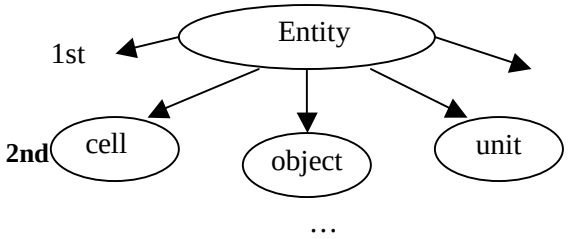

Figure 1. Example of Tree Expansion for Nouns

This form of data presentation makes inspection and observation on the hierarchy among nouns more straightforward. After careful and systematic examination, domain assignment is trickled down to each synset level by level. The same task is performed up to the fifth level. A tree traversal program is executed to trace down the hyponyms and assign domain-tag based on its hypernyms.

1.5 Relational Expansion of Other Parts of Speech

The hierarchy expansion method based on taxonomy mainly applies to nouns. For modifiers such as adjectives and adverbs this general observation does not produce a satisfactory result since "[t]he semantic organization of modifiers in WordNet is unlike…the tree structures created by hyponymy for nouns or troponymy for verbs." [1] However since adverbs/adjectives are often morphologically derived from other major categories, such information can be used to infer domain classification. For example, the adjective 'stellar' is derived directly from the noun 'star.' The term, 'star' is mostly mentioned in an astronomical context. Based on this relation, since 'star' is labeled with 'astronomy' based on Lexical Hierarchy, the adjective 'stellar' can be assigned with the same domain. We combine the tables on the left side and right side of Table 2 Lexical Relation Table to obtain a table organized as follows:

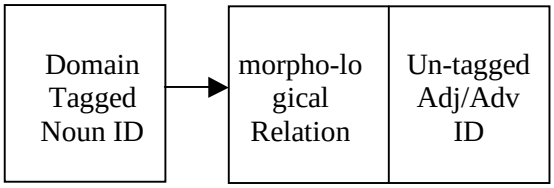

Figure 2. JOIN Method

Later the recordsets that have the relation symbol as "\"(denoted "derived from," refer to Table 2) are extracted and these derived adjectives and adverbs are further assigned with the same domain as the nouns they are derived from.

Results

There are 99,642 unique senses organized by WordNet. By expanding each specific vocabulary coupling with its specific senses, the number of these "word & sense" unique pairs total up to 173,941, which is the basis for all the results.

Parts of Speech	Percentage in Total
Noun	66.87 %
Adjective	17.18 %
Verb	12.69 %
Adverb	3.27 %

Table 5. Percentage of Each Part of Speech in The 173,941 "Word & Senses Pairs" Entries

1.6 Far East Dictionary

There are 20,126 senses that have been assigned with a domain tag with Far East Dictionary, which account for 20.20 % of the total senses (99,642 in total in WordNet). However after expanding it to its synset the total 'word & sense' pairs, there are 42,643 entries being tagged, which account for 24.52 % of the 173,941 pairs in total.

Parts of Speech	Number Tagged	Synset Coverage
Noun	29,946	17.22 %
Adjective	6,188	3.56 %
Verb	6,160	3.54 %
Adverb	349	0.20 %

Table 6. Coverage by POS

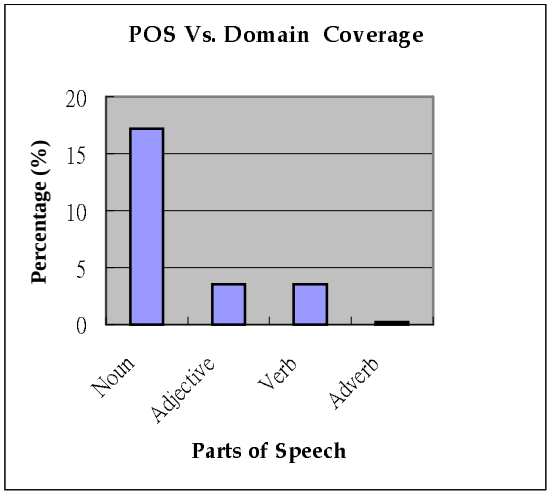

Figure 3. Coverage with Far East Dictionary

1.7 Information Provided by WordNet

The tagging coverage by extracting information directly from WordNet is as follows :

Parts of Speech	Number Tagged	Percentage in Total
Noun	1,826	1.050 %
Adjective	1,501	0.863 %
Verb	2	0.001 %
Adverb	109	0.063 %

Table 7 Coverage with WordNet Info

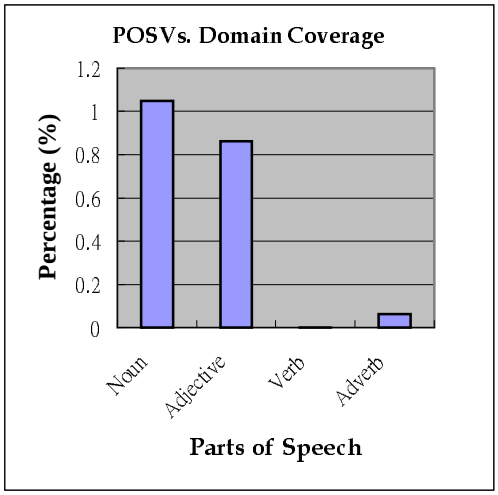

Figure 4. Domain Coverage w/ WN Info

1.8 Domain Tagging Inferred by Lexical Semantic Relation

The result of using the Lexical Relational Structure method is as follows:

Number of Sense Tagged (single level)	Sense Tagged After Tree Expansion	Word & Sense Pairs After Tree Expansion	Percentage In Terms of Total 173,941 Pairs
458	21,781	41,770	24.01%

Table 8. Coverage by Relational

1.9 Tagging by Inheritance through the Lexical Hierarchy of Nouns

We observe the sense meaning of each synset and label the domain by inspection. At first we observe the second level, label the recognizable domain and leave out the ones that are ambiguous. Next we expand to the third level and label the domains. The same procedure is iterated until the hierarchy is expanded to the fifth level. The following is the number of senses that are tagged by inspection and by tree expansion. The total distinct word-sense pairs that have been tagged using 3.4 Taxonomical Method and 3.5 Hierarchical Method is 88,971, which accounts for 51.15% of the total.

Method	Sense Tagged by Inspection	Sense Tagged by Tree Expansion
2nd Level	6	91
3rd Level	292	12,544
4th Level	1,171	28,178
5th Level	373	6,140

Table 9. Tagging Percentage By Inheritance (based on the total of 99,642 senses)

After mapping each sense with all the words in the synset, the result is as follows :

Method	Sense Tagged After Expansion	Percentage In Terms of Total 173,941 Pairs
2nd Level	144	0.08 %
3rd Level	22,478	12.92 %
4th Level	51,607	29.67 %
5th Level	9,707	5.58 %

Table 10. Tagging Percentage By Hierarchy (in the total of 173,942 pairs)

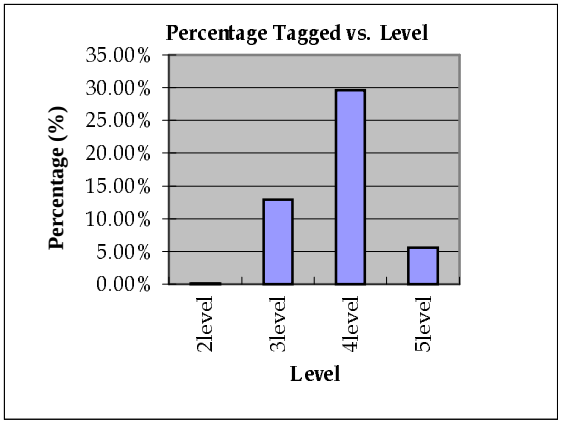

Figure 5. Tagging Percentage By Hierarchy (in the total of 173,942 pairs)

1.10 Relational Expansion of The Modifiers

First we use Table 3 Relations and Pointer Symbols and map it onto the 88,971(51.15% of the total) entries we produced with Method 2.3 & 2.4. Next we extract the rows that contain the symbol "/" which denotes "derived from" to further extend the domain tags from nouns to the modifiers - the adjectives and adverbs. The result is as follows :

Sense Entries with "/"	Expansion to Unique Word & Sense Pairs	Percentage In Total of 173,941 Pairs
2,625	3,452	1.98 %

Table 11. Tagging Percentage of Relational Expansion

Testing and Discussion

The principal testing method we adopt is to first select 200 " word & sense" pairs randomly from the pool of individual results produced by each single method. Method 2.3 is combined with method 2.4; together, they are called the tree expansion method in the following analysis. From Table 12 it is clear that 2.2 Information from WordNet method has the greatest accuracy while 2.1 Far East Dictionary method is ranked second, 2.3 & 2.4 Tree Expansion method placed third, and 2.5 Derivation method is rated last.

Method / Rating	Far East	Word Net	Tree Expansion	Derivati-on
Wrong	18.00%	2.00%	27.00%	24.00%
Acceptable	11.00%	5.50%	7.00%	34.00%
Accurate	71.00%	92.50%	66.00%	42.00%

Table 12. The Accuracy Rating of the Four Methods

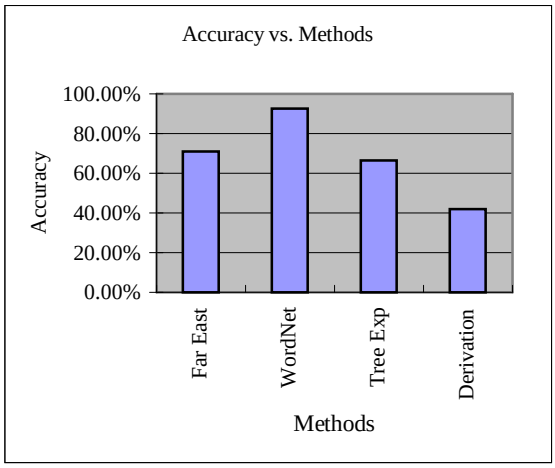

Figure 6. Accuracy vs. Methods

Far East	Word Net	Tree Expansion	Derivation
24.52%	1.98%	51.15%	1.98%

Table 13. Tagging Percentage In The Total 173,942 "Word and Sense" Pairs.

As shown in Figure 7 the tagging entries may overlap. In terms of the accuracy, 2.2 WordNet method should be considered as the best approach, with 92.50% accuracy. This direct information extraction method from WordNet itself does not attain 100% is due to the fact that only certain words in one synset are used in specialized area of studies. For example, in the study of botany, there are a number of terms which indicate the same species, however, only a certain words are the actual scientific names while the rest are merely common names. In our project, our primary objective is to favour the words that belong to the specific area of studies, which is also the main concept upon which our lexical taxonomy is organized.

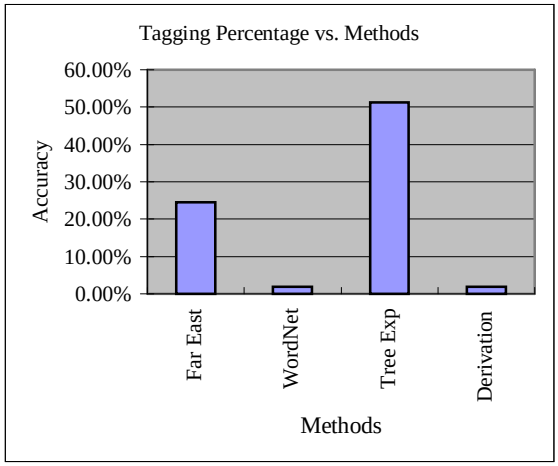

Figure 7. Domain-Tagging Coverage vs. Methods

Based on the extent of domain assignment and the amount of entries covered, Tree Expansion is the most ideal method, with 51.15 % coverage. Both WordNet and Tree Expansion methods have their own disadvantages and advantages, such as time consumption and the extent of coverage. In terms of the WordNet method, extracting data directly from the digital sources is very efficient and the result is more reliable. With high accuracy, the revision that may follow later on in the future would be more straightforward. However, in terms of the extent of coverage, Tree Expansion is still a more effective method. Its result is very encouraging because it contributes to over 51% among the entire domain assignment, with a total of 74% correct or acceptable rate. However, it is worth noticing that for all the entries in WordNet, not every single entry is supposed to be grouped or defined within a specific domain. For instance, all the common grammatical words (a, the, is, etc.) and the high frequency words (hit, kick, smile, etc.) would not and should not belong to a special domain. Although we do not have a realistic measure for recall, the slightly less than 49% coverage of all senses is quite acceptable. So far the number of distinct entries that have been tagged is 103,709, which covers up to 59.62 % of the whole 173,942 word and sense pairs.

2 Future Goals and Improvements

At present our domain tag assignment is still at a preliminary stage, which requires further modifications and improvements. Other method such as bottom up tree traversal is more likely to give rise to a better result with higher accuracy. For example, for a hyponym which falls into the domain of botany, the hypernym is very likely to belong to the domain "biology." Extracting sources from a large corpora grouped by topics is also a reliable approach. For instance, in a journal related to the study in physics, most of the special field-related terms are likely to appear more frequently than in other ordinary sources. Other than extracting information from WordNet itself, other thesauruses in digital files can be taken into consideration as well.

There are a significant number of possible applications that can be contributed by domain tag assignment. Due to the fact that English WordNet is the most fundamental structure upon which a wordnet in other language is based, assigning domain tags to WordNet itself can indeed be expanded to other inter-linked wordnets such as EuroWordNet. By categorizing lexicon into groups of different domains, it will benefit the study of computational linguistics: "word sense disambiguation methods could profit from these richer ontologies, and improve word sense disambiguation performance." [2] Last, but not the least, domain tagging is can be the first realistic step of enriching the linguistic ontology of wordnets so that they can be linked to real-world knowledge and serve as bona fide semantic network for general purpose knowledge processing.

Reference

[1] Christiane Fellbaum. *WordNet : An Electronic Lexical Database.* The MIT Press. Cambridge, Massachusetts, 1998.

[2] Agirre, Eneko et al. *Enriching WordNet concepts with topic signatures.* Proceedings of the NAACL workshop on WordNet and Other lexical Resources: Applications, Extensions and Customizations. Pittsburg, 2001.

[3] Bernardo, Magnini and Gabrela, Cavaglia. *Integrating Subject Field Codes into WordNet.* Proceedings of the LREC conference, 2000.

[4] Chu-Ren Huang, I-Ju E. Tseng, Dylan B.S. Tsai. Translating Lexical Semantic Relations: The First Step Towards Multilingual Wordnets. Proceedings of the 2002 SemaNet Workshop [this volume]. 2002.