Towards a Conceptual Representation of Lexical Meaning in WordNet

Jen-nan Chen
Department of Information Management,
Ming Chuan University
Taipei, Taiwan
jnchen@mcu.edu.tw

Sue J. Ker
Department of Computer Science,
Soochow University
Taipei, Taiwan
ksj@cis.scu.edu.tw

Abstract

Knowledge acquisition is an essential and intractable task for almost every natural language processing study. To date, corpus approach is a primary means for acquiring lexical-level semantic knowledge, but it has the problem of knowledge insufficiency when training on various kinds of corpora. This paper is concerned with the issue of how to acquire and represent conceptual knowledge explicitly from lexical definitions and their semantic network within a machine-readable dictionary. Some information retrieval techniques are applied to link between lexical senses in WordNet and conceptual ones in Roget's categories. Our experimental results report an overall accuracy of 85.25% (87, 78, 89, and 87% for nouns, verbs, adjectives and adverbs, respectively) when evaluating on polysemous words discussed in previous literature.

1 Introduction

Knowledge representation has traditionally been thought of as the heart of various applications of natural language processing. Anyone who has built a natural language processing system has had to tackle the problem of representing its knowledge of the world. One of the applications for knowledge representation is word sense disambiguation (WSD) for unrestricted text, which is one of the major problems in analyzing sentences.

In the past, the substantial literature on WSD has concentrated on statistical approaches to analyzing and extracting information from corpora (Gale et al. 1992; Yarowsky 1992, 1995; Dagan and Itai 1994; Luk 1995; Ng and Lee 1996). However, WSD knowledge acquired from specialized corpora such as various encyclopaedia, book collections or newspaper archive may be biased or incomplete. Words in knowledge are simply partitioned, depending on their use through out the corpus. Those sense partitions do not correspond with those provided in dictionaries. For instance, this type of knowledge would fail to acquire the fish-sense of *bass* from the Wall Street Journal, in the situation where such corpora was employed as a learning resource. Although statistical knowledge acquired from a very large corpus has shown effectiveness in disambiguating text in the same domain, no corpus provides sufficient information to disambiguate unrestricted texts.

Dictionaries provide a ready source of knowledge about senses such as morphology, syntax, definition, example sentences, and collocation. Many references have shown that information in machine-readable dictionary (MRD) is an unbiased knowledge source for WSD (Guthrie et al. 1991; Slator 1991; Li et al. 1995; Chen and Chang 1998b). When a dictionary is directly used as semantic knowledge for any applications of natural language processing, it will lead to immense parameter space. In addition, it is also difficult to master the related semantics for each word sense in a dictionary.

A thesaurus provides conceptual classifications for word senses. It seems to reduce the semantic parameter space, but its lexical and semantic gaps cause another problem, one of incomplete knowledge. For instance, there are ten distinct nominal senses for the word *bank* in WordNet, but only six categories, such as Obliquity(217), Land(342), Store(636), Bare(663), Defence(717) and

Treasure(802), are listed in the Roget's 1911 thesaurus. The Roget's Thesaurus arranges words in a 3-layer hierarchy and organizes over 30,000 distinct words into some 1,000 categories on the bottom layer. These categories are divided into 39 middle-layer sections that are further organized as 6 top-layer classes. Each category is given a 3-digit reference code. However, there is no appropriate Roget's category for the PILE-sense of word *bank* in WordNet. Chen and Chang (1998a) exploited a two-stage approach to fill the sense gap of LLOCE and generate conceptual topics in LLOCE from each sense definition of nominal words in LDOCE. Also, Chen and Chang (1998b) applied these conceptual semantics to disambiguate word senses in the Brown corpus and the Wall Street Journal. They reported conceptual representation acquired from MRD for a word sense did improve the precision of sense tagging on ambiguous words when compared with the corpus-based approach. However, the number of possible senses allowed by thesaurus senses seems very small (only 129 topics). Also, the coverage of their experiment is less impressive since there are over 23,000 dictionary senses for over 16,000 words.

WordNet (Miller 1995; Fellbaum 1998) is a popular on-line lexical reference system for English, organized as a semantic net. Its design is inspired by current psycholinguistic theories of human lexical memory. WordNet (Version 1.6) contains some 118,000 words that are divided into four categories including English nouns, verbs, adjectives and adverbs. Word meanings for each of these categories are organized into sets of synonyms (synset), each representing one underlying lexical concept, and are logically grouped such that words in the same synonym set are interchangeable in some contexts. WordNet contains both individual words and collocations (such as "fountain pen" and "take in"). Different semantic relations, such as hypernymy, hyponymy, meronymy, holonymy, antonymy etc., link the synonym sets. Although it has good coverage (Farreres et al.1998; Kwong 1998) and its synset is much like a thesaurus, the synset fails to provide an explicit classification.

The objective of this paper is to present an automated mapping of dictionary-defined word senses in WordNet into the coarse-grained thesaurus classes in Roget's. The characteristic of this representation is to reduce the lexical dimension and enrich its conceptual information for a lexical word sense. The technique uses an information retrieval approach for extracting conceptual information from available semantic relations for each sense definition in WordNet, such as synsets, hypernym, hyponym etc. To this end, category information in the Roget's is exploited to represent conceptual semantics of word sense in order to characterize the typical context of the sense in question. We are interested in the semantics of four distinctive parts-of-speech, i.e. noun, verb, adjective, and adverb. Applications of this knowledge feature include word sense disambiguation and its related tasks such as information retrieval, machine translation, document classification, and text summarization.

2 Linking WordNet to Roget's

In this section, we apply an information retrieval technique to link MRD senses to thesaurus categories. The current implementation of this approach uses the category information in Roget's to represent conceptual knowledge for WordNet senses. In the following subsections we describe how that is done.

2.1 Mapping Lexical Sense to Conceptual Categories

At first, we treat sense definitions given in WordNet as a raw document in information retrieval. Then, each document may be extended by its synset, hypernym, or hyponym if its description is too vague. For instance, it seems difficult to automatically comprehend the major meaning of word senses from the following list of sense definitions. Consequently, it would also be difficult to generate appropriate representation of conceptual categories for the WordNet senses.

- action.n.1: something done
- issue.n.4: some situation or event that is thought about
- land.adj.1: relating to or characteristic of or occurring on land
- overall.adj.2: including everything

- hard.adv.7: into a solid condition
- hard.adv.10: all the way
- come.v.7: come forth
- make.v.3: make or cause to be or to become

To resolve this problem, we augment the contents of sense definition by some of its related semantic relations in WordNet, if any exist. Table 1 summarises the semantic relations considered for each POS in this paper. In hypernym relation, we only extract words in the immediate parent word's node of a given sense word. In hyponym relation, we extract words in the immediate child of a given sense word.

Table 1 Semantic relations adopted in our paper

	Noun	Verb	Adjective	Adverb
Definition	×	×	×	×
Synonym	×	×	×	×
Hypernym	×	×		
Hyponym	×	×		

With its definition and its semantic relations cast as a document *D* in an IR task, a wealth of IR techniques can be utilized including stopword removal and term weighting (Baeza-Yates and Ribeiro-Neto 1999). Although lexical words appearing in *D* might provide many informative clues for locating categories of a sense, a few of these words, such as function words, are marginally relevant to the sense. To demonstrate this observation, four examples of distinct types of definitions from WordNet are given in Table 2. We find the remaining content words in each document that still characterise the sense of the headword definition while we ignore those stopwords denoted by italics from each of the definitions.

Table 2 Function words marked with italics in sense definition that are marginal to word senses.

Word	Parts of speech	WordNet definition
issue	noun	*one of a* series published periodically
fire	verb	Provide *with* fuel
hot	adjective	*Used of* physical heat
hard	adverb	*with* pain *or* distress *or* bitterness

First, we start with two pre-processing steps including part-of-speech (POS) tagging and stemming for each words in the documents, which is necessary for our conceptual mining algorithm to obtain a good result. The pre-processing steps are done via email to tagger@clg.bham.ac.uk. There are two objectives of this tagging processing on each document: to help prune the irrelevant words in context and to lessen inappropriate categories for content words during the mapping course. After each document is syntactically tagged, the non-information-bearing words can be easily removed from the document. Those pruned words include light verbs, pronouns, determiners, prepositions and conjunctions. Then content words in *D* are represented as a list of keyword-POS pairs, KEY_{D}. Next, for each word *w* in KEY_{D}, we Look up *w* in Roget's to obtain $TOPIC_{W}$. And the set of $TOPIC_{W}$ forms a conceptual-document *CD* of the document *D*. Although this mapping approach is simple, it does introduce a set of noise categories to *CD* for ambiguous words in a document. To remedy this problem, the frequencies of ambiguous words are distributed equally to each of their categories. Thus, in each of these *CD*s, each term is associated with a weighted term frequency (*wtf*) and document frequency (*df*). For instance, there are six categories for lexical word *star* in Roget's. Thus, weighted frequency associated with each of these categories for ambiguous word *star* is assigned to 1/6. Let wtf_{ij} represent the frequency of term t_j in document CD_i, and df_j represent the number of *CD*s where term t_j appears. The relevancy of term t_j to the document CD_i is therefore given by the following weight formula:

$$W_{ij} = C \times wtf_{ij} \times \log(N / df_j),$$

in which N is the number of documents in the collection,

$$wtf_{ij} = \sum_{\{w | w \in KEY_D, t_j \in TOPIC_w\}} \frac{1}{|TOPIC_w|}, \text{ for all } t_j \in CD, \text{ and}$$

C is defined as follows:

C[1] = c_1, if t_j in the definition or same synset with target word,
c_2, if t_j in the immediate hypernym of target word, and
c_3, if t_j in the immediate hyponym of target word and other types of relations.

Those t_j's and their associated weights form a conceptual list for a word sense. We sum up the above description and outline the procedure for assigning Roget's categories to WordNet synset as follows:

Algorithm: Preliminary Linking WordNet to Roget's

Step 1: Given a WordNet synset, merge its sense definition and semantic relations as a document D.
Step 2: Tag each word in D with POS information.
Step 3: Remove all stopwords in D to obtain a list of keyword-POS pairs, KEY_D.
Step 4: Look up w in Roget's to obtain $TOPIC_W$ for all $w \in KEY_D$.
Step 5: Form a conceptual document $CD = \{ t \mid t \in \cup\, TOPIC_W, \text{for all } w \in KEY_D \}$.
Step 6: Compute weighted term frequency and document frequency for all $t \in CD$.

2.2 Illustrated Example: Preliminary Linking WordNet to Roget's

In this subsection, we give an example to illustrate how our approach works to establish preliminary linkage between WordNet and Roget's. Consider a nominal sense definition *star* and its semantic network in WordNet, including synset, and immediate hypernym and hyponym, like the following:

<definition> an actor who plays a principal role
<synset> principal, lead
< hypernym>=> actor, histrion, player, thespian, role player -- (a theatrical performer)

< hyponym> => co-star -- (one of two actors who are given equal status as stars in a play or film)
=> film star, movie star -- (a star who plays leading roles in the cinema)
=> idol, matinee idol -- (someone who is adored blindly and excessively)
=> television star, TV star -- (a star in a television show)

Applying the above algorithm to this example, we have:

Step 1-2: POS_D = { an/ DT, actor/NN, who/WP, play/VBZ, a/DT, principal/JJ, role/NN, principal/NN, lead/NN, role/NN, player/NN, a/DT, theatrical/JJ, performer/NN, co-star/NN, one/ CD, of /IN, two/ CD, actor/NNS, who/WP, be/BER, give/VBN, equal/JJ, status/NN, as/IN, star/NNS, in/IN, a/DT, play/VB, or/CC, film/NN, film/NN, star/NN, movie/NN, star/NN, a/ DT, star/NN, who/WP, play/VBZ, lead/VBG, role/NNS, in/IN, the/DT, cinema/NN, ...}
Step 3: KEY_D= { actor/NN, play/VBZ, principal/JJ, role/NN, principal/NN, lead/NN, role/NN, player/NN, theatrical/JJ, performer/NN, co-star/NN, actor/NNS, equal/JJ, status/NN, star/NNS, play/VBZ, lead/VBG, role/NNS, cinema/NN, ...}

[1] For simplicity, the parameters c_1, c_2 and c_3 are set to 1, 2 and 0.5, respectively, in our experiment.

Step 4: Using Roget's as conceptual representation for each word in KEY_D, we have $TOPIC_{actor}$ ={548, 599, 680, 690, 855}, $TOPIC_{play}$ ={170, 314, 416, 554, 599, 677, 680, 784}, $TOPIC_{principal}$ ={694}, $TOPIC_{lead}$ ={33, 208, 319, 737a}, ...

Step 5: The preliminarily conceptual representation of document D is as follows:

$$CD = TOPIC_{actor} \cup TOPIC_{play} \cup TOPIC_{principal} \cup TOPIC_{lead} \cup \ldots$$

Step 6: For each topic in CD, we have:

$W_{\text{PERFORMER-star}, 599} = 12.48$, $W_{\text{PERFORMER-star}, 694} = 9.12$, ...

The preliminary ranked list of conceptual representation for the PERFORMER-sense of *star* is listed in Table 3.

Table 3 A List of Preliminary Categories for the PERFORMER-sense of *star*.

Roget's Category	Weight	Roget's Category	Weight
The Drama(599)	12.48	Ostentation(882)	0.85
Director(694)	9.12	Precedence(62)	0.82
Importance(642)	4.64	Semitransparency(427)	0.74
Musician(416)	4.60	Direction(693)	0.63
Gravity(319)	2.64	Layer(204)	0.48
Superiority(33)	2.31	Right(922)	0.43
Depth(208)	1.80	Symmetry(242)	0.38
Government(737a)	1.41	Equality(27)	0.34
Worship(990)	1.20	Term(71)	0.32
Love(897)	1.14	Circumstance(8)	0.29
News(532)	1.01	Manifestation(525)	0.29
Business(625)	0.99	Situation(183)	0.27
Conduct(692)	0.99	Fashion(852)	0.25
Plan(626)	0.97	Appearance(448)	0.23
Tendency(176)	0.88	Repute(873)	0.23
Affectation(855)	0.87	...	...

3 Disambiguating Conceptual List

When observing the conceptual list for a lexical sense in Table 3, we find it consists of some irrelevant categories. For instance, when viewing this table, categories Drama, Affectation and Ostentation contributed mainly from the ambiguous word *theatrical* are three disjoint categories in Roget's. Hence, it is inappropriate for all of these categories to appear simultaneously in a conceptual list, as we do not disambiguate the word senses duririg the phase of linking MRD to thesaurus. Thus, a further step of selecting a proper category from this category set is necessary. As Drama has higher score than any other categories on the conceptual list, we may conjecture Affectation and Ostentation belong to be a less relevant category in the list. In this case, it seems reasonable to delete less relevant categories from the list. Following this subsection, we will introduce a method to wipe inappropriate categories off the conceptual list.

3.1 Disambiguating Method

We sum up the above descriptions and outline the identification of relevant conceptual list algorithm as follows.

```
Algorithm: Identification_Relevant_Conceptual_List

Step 1:  While current weighted conceptual list (WCL) is not empty.
         {
Step 2:      Select maximum-scored category (MC) from WCL.
Step 3:      Write MC to relevant conceptual list RCL.
Step 4:   Find implicit lexical word set $W_{MC}$ that contributed to $MC$.
Step 5:   Look up thesaurus and retrieve set of categories $C_W$ for each word w in $W_{MC}$.
Step 6:   Update $WCL$ to $WCL - \bigcup_{x \in W_{MC}} C_x$.
        }
Step 7: Return relevant conceptual list $RCL$.
```

3.2 Illustrated Example: Finding Relevant Conceptual List

In this subsection, we give an example to illustrate how our approach works to find relevant conceptual list from the preliminary linkage. Consider the example shown in Table 3, the conceptual list of PERFORMER-sense of *star* in WordNet.

Step 1: WCL={599(12.48), 694(9.12), 642(4.64), 416(4.60), 319(2.64), 33(2.31), 208(1.80), 737a(1.41), 990(1.20), 897(1.14), 532(1.01), 625(0.99), 692(0.99), 626(0.97), 176(0.88), 855(0.87), …}≠∅
Step 2: MC=599
Step 3: RCL={599(12.48) }
Step 4: W_{599}={ theatrical, role, cinema, thespian }
Step 5: $C_{\text{theatrical}}$={599, 855, 882}, C_{role}={599, 625, 626, 692}, C_{cinema}={599 }, C_{thespian}={599 }
Step 6: After updating WCL, we get
$WCL= WCL- C_{\text{theatrical}} - C_{\text{role}} - C_{\text{cinema}} - C_{\text{thespian}}$
={694(9.12), 642(4.64), 416(4.60), 319(2.64), 33(2.31), 208(1.80), 737a(1.41), 990(1.20), 897(1.14), 532(1.01), 176(0.88), …}

Next we repeat the steps 1-6 until WCL=∅, we can yield the relevant conceptual representation for *star* with PERFORMER sense listed as follows:

The Drama(1.00)[2],
Director(0.79),
Gravity(0.23).

Table 4 Lexical words that contributed to the conceptual list of PERFORMER-sense of *star*.

Lexical word	Category set in Roget's
theatrical, role, cinema, thespian	The Drama(599)
principal	Director(694)
lead	Gravity(319)

Table 5 Relevant conceptual representation for *star* with PERFORMER sense.

(1)	(2)	(3)	(4)	(5)	(6)
Roget's Category	Definition + Synset	(2) + Hypernym	(2) + Hyponym	(2) + Hypernym+ Hyponym	(2) + all Semantic Network
Business	0.12	-	-	-	-
Director	1.00	0.79	1.00	0.79	0.79
Gravity	0.27	0.23	0.27	0.23	0.23
Manifestation	-	-	0.03	-	-
News	-	-	0.11	-	-
Right	-	-	0.04	-	-
Semitransparency	-	-	0.08	-	-
Term	-	-	0.03	-	-
The Drama	-	1.00	0.21	1.00	1.00
Worship	-	-	0.14	-	-

[2] The weigt associated with the most relevant conceptual representation is normalized to 1.

Table 4 shows a summary of the lexical words that contributed to each category in the conceptual list. And table 5 presents results of the relevant conceptual list assigned by our system for the PERFORMER-sense of *star* when definition is mixed with its various semantic relations. For instance, the second column shows the results for relevant conceptual representation when definition and synset are applied to our algorithms. And the fifth column shows three most relevant topics are extracted from the original topical list.

4 Experimental Results

We have conducted some preliminary experiments on this approach, by running tests on all senses of WordNet. To evaluate the algorithm, we selected the 35 most polysemous words used in recent WSD experiments (Yarowsky 1992; Luk 1995; Leacock and Chodorow 1998) from the test set. These selected words used in the evaluation consist of much more difficult words, in which degrees of ambiguity are over average (see Table 6). These highly polysemous words include 20 nominal words, 5 verbal words, 6 adjectival, and 4 adverbial words. These 20 nominal words are *bank* (10), *bass* (8), *bow* (9), *cone* (4), *crane* (2), *duty* (3), *galley* (4), *interest* (7), *issue* (11), *jack* (11), *mole* (6), *poll* (5), *port* (5), *sentence* (3), *slug* (3), *space* (10), *star* (7), *suit* (5), *table* (6), and *taste* (7). The five verbal words include: *request* (3), *order* (9), *play* (29), *lie* (8), and *fire* (9). The six adjectival words are *hard* (11), *blue* (11), *cool* (12), *rich* (12), *easy* (15), and *heavy* (29), while the four adverbial words include: *hard* (10), *right* (14), *flat* (6), and *short* (8). The numbers in parenthesis followed by lexical words denote the degree of ambiguity of the words. The results show that, on the average, our conceptual knowledge mining algorithm presented correctly 87%, 78%, 89%, and 87% of categories for each sense in the nominal, verbal, adjective, and adverb words in WordNet. Tables 7(a)-7(d) show the conceptual representation of four typical words selected from our experiment in which each polysemous word comes from a distinct category in WordNet. It shows that the wrong conceptual components in a conceptual list have a very low ratio. To our knowledge, there is no current method that attempts to identify automatically the conceptual senses of all words in MRD, so we can't make a practical comparison.

Table 6 Average mapping results for the appropriateness of knowledge representation.

POS	Average ambiguity		Average precision of concepts on relevant category[3]
	Whole set	Test set	
noun	2.73	6.3	87%
verb	3.57	11.6	78%
adjective	2.80	15.0	89%
adverb	2.50	9.5	87%

[3] This ratio is defined by the number of relevant concepts divided by the number of possible concepts for a given sense.

Table 7(a) An example of conceptual representation for senses of an ambiguous word *taste* in WordNet

Target word	POS	Sense definition	Conceptual representation
taste	noun	distinguishing a taste by means of the taste buds	Taste(1.00), Idea(0.24), Intelligence(0.16), Effect(0.07)
		the faculty of taste	Taste(1.00), State(0.41), Motive(0.28), Intelligence(0.16), Will(0.09), Skill(0.08)
		the sensation that results when taste buds in the tongue and throat convey information about the chemical composition of a soluble stimulus	Taste(1.00), Saltiness(0.27), Wonder(0.26), Simpleness(0.24), Motive(0.22), Sweetness(0.22), Excitation(0.21), Sourness(0.19), Airpipe(0.18), Acridity(0.17), Liquefaction(0.15), Irascibility(0.12), Effect(0.11), Transfer(0.08), Vehicle(0.07), Wit(0.07), Remedy(0.06), Food(0.04)
		delicate discrimination (especially of aesthetic values)	Taste(1.00), Motive(0.37), Fashion(0.32), Conduct(0.16), Penalty(0.16), Caution(0.13), Sociality(0.06), Party(0.04), Leisure(0.04), Ugliness(0.03)
		a brief experience of something	Eventuality(1.00), Taste(0.67)
		a strong liking	Desire(1.00), Pleasure(0.43), Taste(0.24), Touch(0.13), Irresolution(0.08)
		a small amount eaten or drunk	Quantity(1.00), Drunkenness(0.67), Taste(0.37), Imperfection(0.36), Acridity(0.28)

Table 7(b) An example of conceptual representation for senses of an ambiguous word *fire* in WordNet.

Target word	POS	Sense definition	Conceptual representation
fire	verb	Bake in a kiln	Furnace(1.00), Calefaction(0.87), Location(0.57), Servant(0.52), Food(0.36)
		destroy by fire	Poverty(1.00), Calefaction(0.74), Evil(0.32)
		cause to go off	Calefaction(1.00)
		go off or discharge	Calefaction(1.00)
		start firing a weapon	Calefaction(1.00), Arms(0.78), Physical Pain(0.21)
		call forth of emotions feelings and responses	Excitation(1.00), Feeling(0.83)
		drive out or away by or as if by fire	Meaning(1.00), Calefaction(0.70), Repulsion(0.54), Hardness(0.53), Haste(0.39)
		provide with fuel	Provision(1.00), Calefaction(0.27), Fuel(0.25)
		terminate the employment of	End(1.00), Stealing(0.27), Calefaction(0.26), Business(0.17)

Table 7(c) An example of conceptual representation for senses of an ambiguous word *cool* in WordNet.

Target word	POS	Sense definition	Conceptual representation
cool	adj	marked by calm self-control especially in trying circumstances	Circumstance(1.00), Dissuasion(0.58)
		feeling or showing no enthusiasm	Feeling(1.00)
		calm and unemotional	Dissuasion(1.00)
		(music) restrained and fluid and marked by intricate harmonic structures often lagging slightly behind the beat	Slowness(1.00), Amorphism(0.54), Form(0.54), Prohibition(0.49), Poetry(0.46), Complexity(0.37)
		(informal of a number or sum) without exaggeration or qualification	Number(1.00)
		(informal) marked by great skill or facility	Skill(1.00), Disinterestedness(0.23)
		neither warm or very cold giving relief from heat	Cold(1.00), Materiality(0.77), Physical Sensibility(0.54), Heat(0.42),Inutility(0.28)
		psychologically cool unfriendly or unresponsive or showing dislike	Feeling(1.00), Friendship(0.89), Agitation(0.57), Interpretation(0.54), Intellect(0.47), Poverty(0.44), Mankind(0.42), Elasticity(0.41), Arrangement(0.31), Indifference(0.22), Dislike(0.21), Enmity(0.11)
		(color) inducing the impression of coolness used especially of greens and blues and violets	Blueness(1.00), Cause(0.21), Hate(0.11)

Table 7(d) An example of conceptual representation for senses of an ambiguous word *hard* in WordNet.

Target word	POS	Sense definition	Conceptual representation
hard	adv	with effort or force or vigor	Exertion(1.00), Strength(0.30), Materiality(0.14), Waste(0.07), Nomenclature(0.06)
		to the full extent possible all the way	Impenitence(1.00), Space(0.79), Loudness(0.62)
		slowly and with difficulty	Difficulty(1.00), Exertion(0.34), Materiality(0.29), Intellect(0.22), Elegance(0.11), Inclusion(0.11), Completion(0.10), Permanence(0.07), Disinterestedness(0.06), Requirement(0.05)
		causing great damage or hardship	Adversity(1.00), Impenitence(0.46), Loss(0.31), Disinterestedness(0.18)
		with firmness	Impenitence(1.00), Perseverance(0.72)
		earnestly or intently	Impenitence(1.00)
		with pain or distress or bitterness	Pain(1.00), Impenitence(0.64)
		very near or close in space or time	Impenitence(1.00), Time(0.82), Closure(0.80)
		into a solid condition	Hardness(1.00), Belief(0.39), Qualification(0.35)
		indulging excessively	Impenitence(1.00), Lenity(0.73), Drunkenness(0.38), Redundancy(0.34), Conduct(0.17)

5 Discussion

5.1 Document Size

From the experiment, we find including synset, hypernym and hyponym into sense definition leads to a positive improvement in the quality of knowledge representation when compared to considering sense definition only. This treatment of the knowledge acquisition is both effective and economic; it takes about 20 minutes on a Compaq 500 to link all words in WordNet to Roget's.

5.2 Collocations

We do not exploit collocation, though we believe this information may useful for improving the quality of the representation. When we observe the result, we find some of the collocations are either recognized or will not influence the ranking of the relevant topics. For instance, in the following definition:

> Blue: used to signify the Union forces in the Civil War who wore blue uniforms.

The collocation *Civil War* is treated as two separated entities in our system. But it does not influence the description of the topics. The resulting topics are listed below:

> Blueness, Agent, Warfare, Government, Clothing, Indication.

However, consider the following definition:

> Issue: come out of.

The collocation *come out* was treated as a two separated entities in this system. It does influence the description of the topics when we do not treat it as a related entity. And we derive the topics ranked by weights as follows:

> Egress, Focus, Posterity, Disease, Completion.

We observed if the collocations consisting of verb+(compound) preposition combinations are excluded from our experiment, they would weaken the major topics and tend to include inappropriate topics in the conceptual representation.

5.3 POS Tagging Error

If a lexical item is mislabeled in regard to part of speech, errors in conceptual representation will understandably follow. For instance, when one of the MOLE senses -- (a small congenital pigmented spot on the skin), and its hypernym --a mark or flaw that spoils the appearance of something (especially on a person's body) was inputted to tagger, the lexical word "spoils" labeled noun instead of verb. Since there is only one Roget's category, booty, for the nominal sense, it was not possible for our approach to select a more appropriate category such as deterioration. However, this kind of error is very limited.

5.4 Applications to Natural Language Processing

One of the main applications for our conceptual representation for a sense is to resolve the issue of word sense disambiguation. Consider the following text selected from the Brown Corpus, which contains a FACTORY sense of the ambiguous word *plant*.

> ...buy a package program from an insurance <u>company</u> simply because it <u>works</u> for another **plant**. But even if that other **plant** <u>employs</u> the same number of <u>workers</u> and makes the same <u>product</u>, there are other facts ...

The sense of *plant* can be disambiguated as FACTORY, since the *company*, *work*, *employ*, *worker*, and *product* have concepts WORKSHOP, PRODUCTION, and FACILITY overlapped with the relevant conceptual list of FACTORY-*plant* sense.

Besides, the proposed conceptual knowledge for a word sense can allow a lexicographer to fill in the gaps, either lexical or semantic, in the thesaurus. For instance, there are three coarse senses for the lexical word *star* in Roget's but a PROFICIENT(= someone who is very highly silled) sense is not acquired from the thesaurus. This sense gap can be successfully filled after running this algorithm.

6 Conclusion

There are many types of knowledge representation ranging from sense number in MRD to topic (category) in thesaurus. In this paper we propose an automatic construction of the appropriate conceptual knowledge to each of the words in WordNet. Each sense of lexical words in a WordNet is represented by a list of relevant concepts in a thesaurus. The list of concepts is regarded as a vector in the multi-dimensional space of topics. This vector representation derived from our approach will provide a backbone for disambiguating the semantics of the applications of natural language processing. The proposed method shows more topical information for a word sense than lexical description in a dictionary and it is easily transferred from any MRD to any thesaurus. The strength of our approach is it does not require specific and substantial corpus to derive semantic knowledge for general-domain texts. We use only existing knowledge sources to aid semantic interpretation and then map it to prespecified categories. However, the specific-domain such as law may be acquired from the general-domain by the adaptive steps on the specific texts. (Chen and Chang 1998b)

Currently, we have applied a set of IR techniques for linking an English/Chinese bilingual dictionary to English WordNet. Future work will focus on constructing an elementary framework of Chinese WordNet from this linking and the experimental results of this paper.

Acknowledgements

This work was partially supported by the NSC, ROC, under grants 89-2213-E-130-006. We would like to thank the anonymous referees for their valuable comments. Any errors that remain are solely the responsibility of the authors.

References

Baeza-Yates, R. and B. Ribeiro-Neto. 1999. *Modern Information Retrieval.* Addison Wesley.

Chen, J.N. and J.S. Chang. 1998a. TopSense: A topical sense clustering method based on information retrieval techniques on machine readable resources. *Special Issue on Word Sense Disambiguation, Computational Linguistics*, 24(1), 61-95.

Chen, J.N. and J.S. Chang. 1998b. A Concept-based Adaptive Approach to Word Sense Disambiguation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, 237-244, Montreal, Canada.

Church, K.W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, 136-143, Austin, Texas, USA.

Dagan, I. and A. Itai. 1994. Word Sense Disambiguation Using a second language monolingual corpus. *Computational Linguistics,* 20(4), 563-596.

Farreres, X., G. Rigau, and H. Rodriguez. 1998. Using WordNet for building WordNets. In *Proceedings of the Usage of WordNet in Natural Language Processing Systems*, Montreal, Canada, 65-72.

Fellbaum, Christiane. 1998. Word-Net: An electronic lexical database. MIT Press, Cambridge, MA.

Gale, W.A., K.W. Church, and D. Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation,* 101-112.

Guthrie, J., L. Guthrie, Y. Wilks, and H. Aidinejad. 1991. Subject-dependent co-occurrence and word sense disambiguation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 146-152.

Kwong, O.Y. 1998. Aligning WordNet with additional lexical resources. In *Proceedings of the Usage of WordNet in Natural Language Processing Systems*, Montreal, Canada, 73-79.

Leacock, C. and M. Chodorow. 1998. Using corpus statistics and WordNet relations for sense identification. *Special Issue on Word Sense Disambiguation, Computational Linguistics*, 24(1), 147-165.

Li, X., S. Szpakowicz and S. Matwin. 1995. A WordNet-based algorithm for word semantic sense disambiguation. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence IJCAL-95,* Montreal, Canada.

Luk, A.K. 1995. Statistical sense disambiguation with relatively small corpora using dictionary definitions. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 181-188.

Miller, G.A. 1995. Word-Net: A lexical database for English. In *Communication of the ACM*, 38(11), 39-41.

Ng, H.T. and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach, In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, 40-47.

Roget's Thesaurus of English words and Phrases. 1911, Longman Group UK Limited.

Slator, B. 1991. Using context for sense preference. In Zernik (ed.) Lexical acquisition: Exploiting on-line resources to build a lexicon, Lawrence Erlbaum, Hillsdale, NJ.

Yarowsky, D. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, 454-460, Nantes, France.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 189-196.