E-HowNet: the Expansion of HowNet

Shu-Ling Huang, You-Shan Chung, and Keh-Jiann Chen

Institute of Information Science, Academia Sinica
{ josieh, yschung, kchen}@iis.sinica.edu.tw

1. Introduction

HowNet is an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents [1]. Each concept is represented and understood by their definition and association links to other concepts. To Compare with WordNet, HowNet's architecture provides richer information apart from hyponymy relations. It also enriches relational links between words via encoded feature relations. The advantages of the HowNet are a) inherent properties of concepts are derived from encoded feature relations in addition to hypernym concepts, and b) information regarding conceptual differences between different concepts and information regarding morph-semantic structure are encoded. HowNet's advantages make it an effective electronic dictionary in NLP area. In recent years, HowNet has been applied to the researches of word similarity calculation[2], machine translation[3], and Information Retrieval[4] etc.

However, what make us interest here is how to use HowNet to achieve mechanical natural language understanding. When we say that a sentence is 'understood', which means that the concepts and the conceptual relationships expressed by the sentence is unambiguously identified and we can make right inferences and/or responses. Similarly, computer understanding required a representational framework which suffices to represent knowledge about lexical concepts and to perform semantic composition and sense disambiguation processes. At present, HowNet has not focused on the aspect of semantic composition. We therefore propose a framework extended from HowNet, called E-HowNet, to deal with this problem. E-HowNet is a frame-based entity-relation model [5] which intends to achieve following goals:

a. Word senses (concepts) can be defined by not only primitives but also any well-defined senses and sense relations.

b. Near canonical representations.

c. Semantic composition and decomposition capabilities.

d. Universal and language independent representation.

In the following section, we will discuss a number of feature expansions of the E-HowNet. In Section 2, we describe using multi-level concept definitions to enhance ontological relations between concepts. Then, in Section 3, in order to achieve semantic composition, we propose a uniform representational framework for both function words and content words. In section 4, *function*, a special kind of relation which maps concepts to concepts, is introduced. We discuss the issues of sense omission and filling semantic gaps by automatic deduction in Section 5. Summarization and conclusion are given in Section 6.

2. Build up ontological relations between concepts by multi-level definitions

HowNet uses the smallest unit of meaning, called sememe, to define concepts. For example, '狗 dog' is defined as def: {livestock|牲畜}.Using primitives to define concepts not only causes information degrading but also fails to establish some important ontological relations between concepts. For example, HowNet defined '獅子狗 Beijing dog' as def: {livestock|牲畜} as well, in which we lose the hyponymy relation toward 'dog'. Thus, similar to HowNet, we adopt entity-relational model to define word sense, except that a concept is defined by simpler or synonym concepts instead of semantic primitives only and all attribute relations are explicitly expressed [6]. In E-HowNet '獅子狗 Beijing dog' is defined as def:{dog|狗:source={Beijing|北京}}. It denotes the ontology relation between 'dog' and 'Beijing dog' by using concept 'dog' as head sense, so that the concept definition itself forms an ontological network.

Nevertheless, the set of HowNet sememes (semantic primitives) are also adopted at E-HowNet for the ground-level definitions. In E-HowNet, new concepts are defined by any well-defined concepts and a definition can be dynamically decomposed into lower level representations until ground-level definition is reached, in which all features in the definitions are sememes. For instance, the top level definition of '文學系 department of literature' is as (1)

(1) def: {school department|學系:
predication={teach|教:
location={~},
theme={literature|文}}}.

Since the concept '學系 school department' is not a primitive concept, the above definition can be further extended into the primitive level definition (1'). The notation of '~', as in HowNet, refers to the head concept of the definition which is 'school department|學系' in (1).

(1') def:{InstitutePlace|場所:
domain={education|教育},
predication={study|學習:
location= {~}},
predication= {teach|教:
location = {~},
theme={literature|文}}}.

Such a multi-level representational framework not only makes sense definitions more precise and readable but also retains the advantage of using semantic primitives to achieve canonical sense representation. In addition, in order to achieve unambiguous definitions and language independent, in E-HowNet, WordNet synsets were adopted as an alternative vocabulary for conceptual indexing and representation. Take (2) as an example,

(2) exhibit as evidence 證物

a. Original E-HowNet definition
def:{physical|物質:
domain={police|警},
telic= {prove|證明:
instrument={~}}}.

b. Definition is in terms of WordNet Synset id-numbers
def:{[00010572N]:
domain={[06093563N]},
telic= {[00686544V+01816870V]:
instrument= {~}}}.

c. Definition is in terms of WordNet Synset concepts
def:{<substance>:
domain={<police>},
telic= {<testify+corroborate>:
instrument={~}}}).

3. Uniform representation for content words and function words for semantic composition

HowNet's objective is to define word sense and to express synonyms with the same representation. In addition, similarity of word senses can be derived by comparing their definitions. HowNet works well for defining content words, but it does not provide a good representational framework for expressing the senses of function words. For instance, the senses of function words have same semantic head of {FuncWord|功能詞:…}. If we aim to perform semantic composition, it is essential to have a uniform framework to express both functional sense and content sense.

The sense of a natural language sentence is composition of the senses of constituents and their relations. Lexical senses are processing units for sense composition. Conventional linguistic theories classify words into content words and function words. Content words denote entities and function words without too much content sense mainly mark grammatical functions. However in fact there is no clear cut distinction between two classes particularly for Chinese language. We can only say that the major sense of a function word denotes relations and with less content senses. For example, '被 by' is a preposition that introduces an agent role/relation without additional content sense. On the other hand, the adverb 'gently' in a sentence bridges a 'manner' relation between its content sense 'gentle' and the action indicated by the sentential head. In contrast, content words, such as verbs and nouns, have more content senses and less (or underspecified) relational senses. A verb denotes an event and also contributes senses of its event roles. A noun refers to objects while plays the roles of verb arguments or modifiers of nouns. Therefore it is clear that all words contain two types of senses, relation sense and content sense. The sense spectrum for syntactic categories is as shown in Table 1. For a lexical knowledge representation system, it is necessary to encode both relation senses and content senses in a uniform framework. E-HowNet is an entity-relation model to achieve representations of content/function word senses and sentence/phrasal senses. Some E-HowNet representations of word senses are shown in Table 2.

Table 1. The sense spectrum for syntactic categories

Function words	Content words
Relational senses ←--→ Content senses	
De, prepositions, conjunctions, adverbs, ……………………, adjectives, verbs, nouns	

Table 2. Examples of E-HowNet representations

Word	POS	Definition
因為	Cb	reason ={ }
下雨	Va	{rain|下雨}
衣服	Na	{clothing|衣物}
都	Da	quantity={complete|整}
濕	Vh	{wet|濕}
了	Ta	aspect={Vachieve|達成}

3.1 Lexical representations and basic semantic composition processes

In E-HowNet, the senses of function words are represented by semantic roles/relations [7]. For example, the conjunction 'because' is defined as shown in (3), which links two entities x and y:

(3) because 因為

def: reason={}; which means reason(x)={y} where x is the dependency head and y is the dependency daughter of '因為'.

In a semantic composition process, if two constituents are syntactically dependent, their E-HowNet representations will be unified according to the following basic composition processes:

If a constituent *B* is a dependency daughter of constituent *A*, i.e. *B* is a modifier or an argument of *A*, then unify the semantic representation of *A* and *B* by the following steps.

Step 1: Disambiguate the senses of *A* and *B*.

Step 2: Identify semantic relation between *A* and *B* to derive relation(*A*)={*B*}.

Step 3: Unify the semantic representation of *A* and *B* by insert relation(*A*)={*B*} as a sub-feature of *A*.

Since the methods for word sense disambiguation and relation identification are out of the scope of this paper. We will not address these two issues here.

In following sentence (4), we'll show how the lexical concepts are combined into the sense representation of the sentence.

(4) Because of raining, clothes are all wet. 因為下雨，衣服都濕了

In the above sentence, '濕 wet', '衣服 clothes' and '下雨 rain' are content words while '都 all', '了 Le' and '因爲 because' are function words. Their E-HowNet sense representations are shown in Table 2. The difference of their representation is that function words start with a relation but content words have under-specified relations. If a content word plays a dependency daughter of a head concept, the relation between the head concept and this content word will be established after parsing process. Suppose that the following dependency structure and semantic relations are derived after parsing the sentence (4).

(5) S(reason:VP(Head:Cb: 因爲 |dummy:VA: 下雨)|theme:NP(Head:Na: 衣服) | quantity: Da:都 | Head:Vh:濕|particle:Ta:了)。

After unification process, the following semantic composition result (6) is derived. The representations of dependency daughters became the feature attributes of the sentential head 'wet|濕'.

(6) def:{wet|濕:
theme={clothing|衣物},
aspect={Vachieve|達成},
quantity={complete|整},
reason={rain|下雨}}.

In (5), function word '因爲 because' links the relation of 'reason' between head concept '濕 wet' and '下雨 rain'. The result of composition is expressed as reason(wet|濕)={rain|下雨}, since for simplicity the dependency head of a relation is normally omitted. Therefore reason(wet|濕)={rain|下雨} is expressed as reason={rain|下雨}; theme(wet|濕)={clothing|衣物} is expressed as theme={clothing|衣物} and so on.

4. Semantic roles and functions

E-HowNet is an entity-relation model as described above, in which entities indicate objects or events which have concrete content sense, and relations link the semantic relations between entities. There are two different types of relations, semantic roles and functions. All semantic roles are binary relations rel(x,y). The parameter x usually is dependency head and we write rel(x,y) as rel(x)={y}, which reads as 'rel of x is y'. For example, Agent(eat)={dog} means 'agent of eating is a dog'. In {eat:

agent={dog}}, ‘agent={dog}’ is an abbreviation of agent(~)={dog} where ~ denotes the head concept which is ‘eat’ in this example. A relation rel(x)={y} is considered as a mapping from domain(x) to range(y). Domain and range are constrained for different relations. In HowNet the range of attribute types of relations is constrained by their attribute-values|屬性值. For instance, the color-values are blue|藍, red|紅, green|綠 and so forth. Other kind of semantic roles are participants of events, such as agent, theme, goal,…, etc.. Their range values are constrained depending of the head events.

Function is a special kind of relation which maps concept/concepts to a specific concept in the same domain. It is not to be used to establish the thematic relation or property attribute between two parameters, but to transform a concept to a new concept. Function has compositional property. New function can be constructed by composition of many functions of the same type. For instances, the kinship function of Father(Father(x)) denotes grandfather of x and the direction function of North(East()) denotes the direction of north-east. Both are the composition functions of basic functions. Function expression is written as rel(x) and treated as a concept or sememe in E-HowNet expression, (7) is a typical example.

(7) vehicle headlight 車燈

def: {PartOf({LandVehicle|車}):
telic={illuminate|照射:
instrument={~}}}.

In the above definition, ‘PartOf' is a function while ‘telic’ and ‘instrument’ are semantic roles. ‘Telic’ is used to build relation between the target object and the event, so as ‘instrument’. On the contrary, ‘part of’ does not build relation but mark the range of the target object.

In E-HowNet, we also regard and-or relation, question and negation relation as functions. Their usage is like following examples:

(8) father-in-law 岳父/公公

def: {father({spouse({x:human|人})})}.

(9) Eastern Taiwan 東台灣

def:{east({Taiwan|台灣})}.

(10) get in and out 進出

def: {or({GoInto|進入},{GoOut|出去})}.

(11) why 為何

def: question({reason({x:event})}).

(12) be distracted 分神

def: {not({attentive|專注})}.

However, semantic role also has the form rel(x) which expresses an underspecified value and expects to fill in a specific value to complete the expression. Examples are as below:

(13) a. 'wavelength' 波長

def: length({波}).

b. 'wavelength 10 km' 波長十公里

def: length({波})={10 公里}.

c. 'electric wave which has 10 km wavelength' 波長十公里的電波

def:{電波:length(~)={10 公里}}.

In order to achieve automatic feature unification processes, we organized relations into a hierarchical structure just as taxonomy for entities. A hyponym relation entails its hypernym relations. The taxonomies of semantic roles are shown as table 3 and 4. The taxonomy of function is shown in table 5. The complete taxonomy of E-HowNet ontology can be seen at http://mt.iis.sinica.edu.tw/~mhbai/taxonomy/.

Table 3. The taxonomy of semantic role for objects.

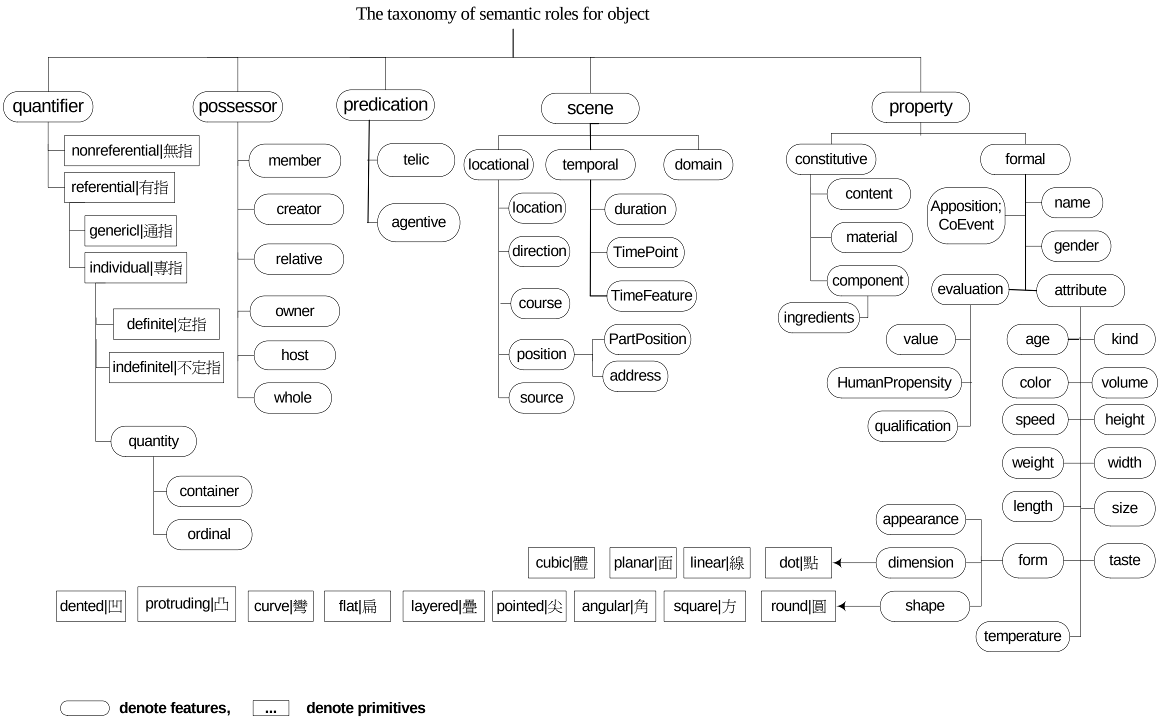


Table 4. The taxonomy of semantic role for events.

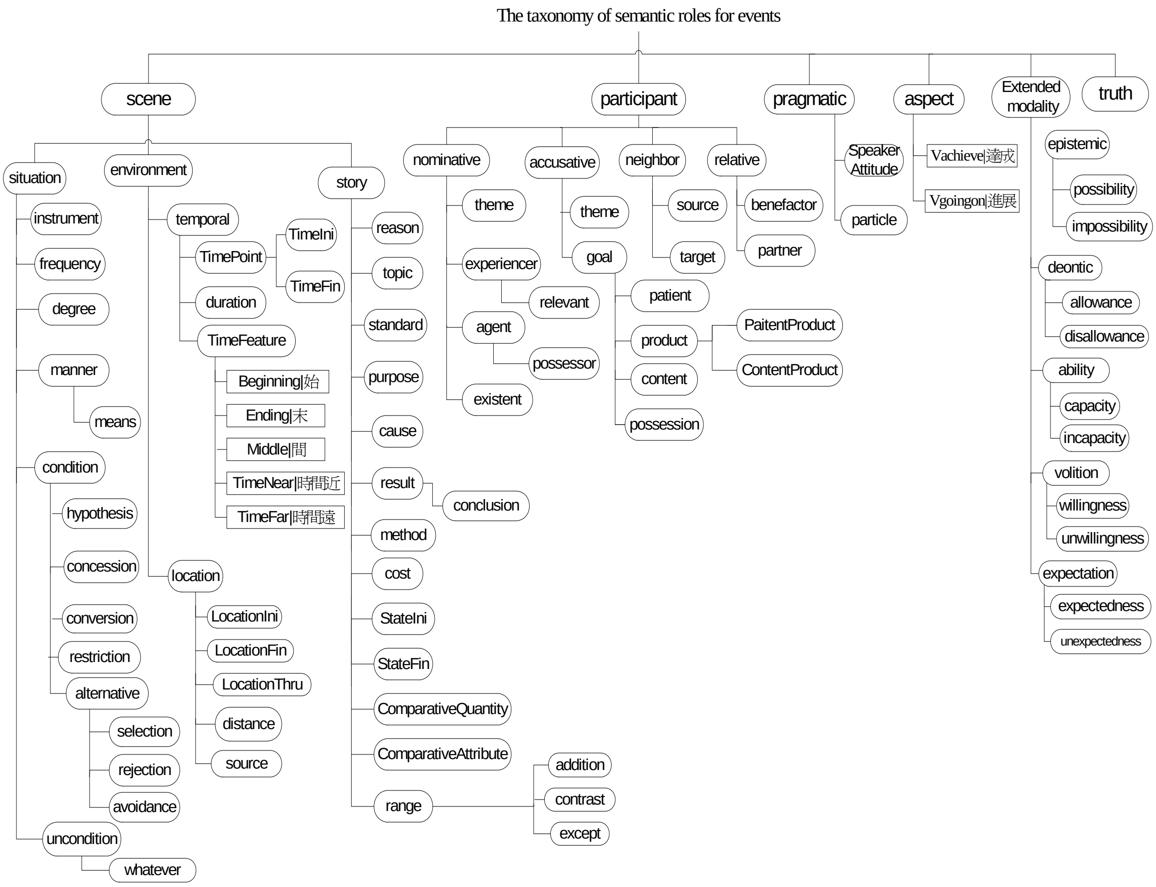
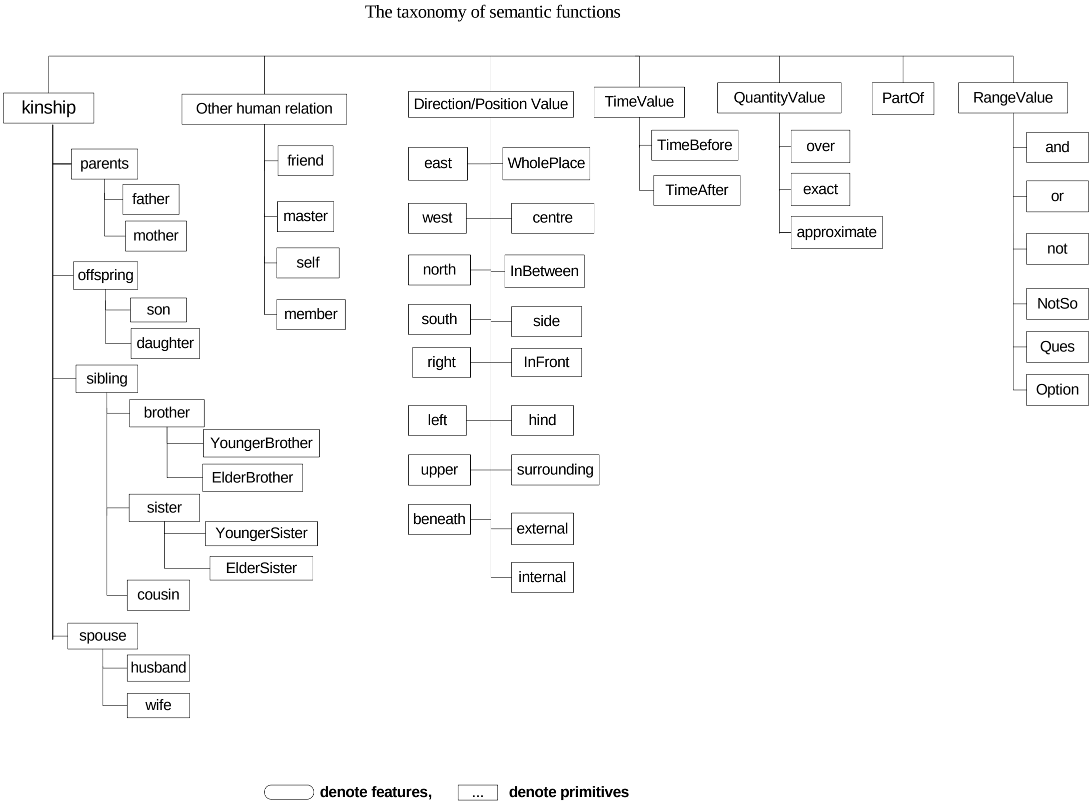


Table 5 The taxonomy of functions



5. Filling semantic gaps by automatic deduction

In real implementations of semantic composition, we have found filling semantic gaps an important task, because some sentence elements are frequently omitted from surface sentences. In order to restore sense omissions, event frames and constriction patterns became an integral part of the E-HowNet system. We have not only established object-attribute relations, but also revealed the participants in an event. For instance, 'color' is a semantic role that builds relation between an object x and its color range y, express as color(x)={y}. In following sentence (14-16), we demonstrate how to restore sense omissions by object-attribute relations.

(14) I like the red? 我喜歡紅的

def:{ FondOf|喜歡:
    agent={I|我},
    target={object|物體:
        color={red|紅}}}.

Because the semantic role 'color' is an attribute of objects, it implies an object was missing in the sentence (14) and thus it is known that the target of 'like' has to be recovered from context. Similarly Quantitative Determinative is a semantic role that establishes the relation between an object and its quantity. In sentence (16), we easily find the object is omitted.

(15) few 少數

def: quantity={few|少}.

(16) There are only a few dare to speak 敢說話的是少數

def:{dare|敢於:
content={express|表示},
experiencer={object|物體:
quantity={few|少}}}.

Another way of filling semantic gaps is referring to the construction meaning and a mapping table to connect the grammatical functions and fine-grained semantic roles [8]. The most typical example is the comparative construction for '比 bi'. The sense of 'bi' comprises a complex argument structure which is shown in (17), and in the following sentence (18), we'll see its implementation:

(17) 'bi'比 =def: contrast={} in the course-grained event frame of {AttributeValue: theme={},contrast={},quantity (or degree)={}, manner={}, location={},time={}}.

(18) I am taller than him by a head.我比他高一個頭

Surface structure: theme[NP]+contrast[PP[比]]+Head[V]+quantity

Parsing result: {tall|高:
theme={I|我},
contrast={he|他},
quantity={one head|一個頭}}.

Through devising a mapping table to connect the grammatical functions and fine-grained semantic roles, we help the machine to find the corresponding constituents in the comparative sentences and fill the semantic gaps. (18') is the result.

Table 6. Mapping table for the fine-grained semantic roles

Fine-grained Semantic Roles	Thematic Roles	Grammatical Functions
Profiled_Item+(Profiled_Attribute)	Theme; Experiencer	Subject
Standard_Item+(Standard_Attribute) Comparison_set	Contrast	Object[PP[bi]]
Attribute_Value	Head	Verb
Degree	Quantity; Degree	Complement
Manner	Manner	Adjunct (Manner)
Place	Location	Adjunct (Location)
Time	Time	Adjunct (Time)

(18') My height is one head taller than his height.我的身高比他的身高高一個頭
def:{tall|高:
Profiled_Item={I|我},
Profiled_Attribute={height|身高},
Standard_Item={he|他},
Standard_Attribute={height|身高},
Degree={one head|一個頭}}.

6. Conclusion and the future work

HowNet proposed a new model to represent lexical knowledge, which inspires us using this framework to achieve the task of natural language understanding by computers. E-HowNet conferred each concept a semantic type, and defined the relation between these types. Hence we have a consistent view to check all concepts, and the computer can understand plain context.

Semantic composition is the key part of computer understanding. In this paper, we design a uniform representation system for both function words and content words to achieve semantic composition. We also add 'functional composition' to extend the expression of new concepts and make the word definition more accurate. Since sense omission will cause misunderstanding, we try to fill the semantic gaps by automatic deduction under the framework of E-HowNet.

To achieve semantic composition, we started with the representations of core lexical senses and proposed a methodology for predicting and deriving sense representations for unknown compounds [9]. The idea is by using the existing definitions of similar compounds, such as ‘搬運工 porter’, to predict, for example, the sense of an unknown compound ‘牧工 hired herdsman’ automatically, and not be confused with similar structure but unrelated compound ‘美工 art designing’. We also implement a sense-representation derivation system for determinative-measure compounds and obtained a good result [10].

However, we are still facing some problems. Apart from sense disambiguation, discordance between syntactic structure and semantic relations is another critical problem. We have to find out the mapping rules to match coarse-grained syntactic arguments to fine-grained semantic relations, in which coercion and gap filling processes are an integral part of the mechanism. They will be addressed more in our future research.

Acknowledgement:

This research was supported in part by the National Science Council under a Center Excellence Grant NSC 96-2752-E-001-001-PAE and Grant NSC96-2221-E-001-009.

References

[1] Zhendong Dong, Qiang Dong, 2006, HowNet and the Computation of Meaning. World Scientific Publishing Co. Pte. Ltd.

[2] 劉群、李素建，2002，基於知網的詞彙相似度計算，第三屆中文詞彙語義學研討論論文集，台北。

[3] 董振東，1999，Bigger Context and Better Understanding -- Expectation on Future MT Technology, Proc. of the International Conference on Machine Translation & Computer Language Information Processing, pp.17-25.

[4] Bonnie J. Dorr, Gina-Anne Levow, Dekang Lin, 2000, Construction of Chinese-English Semantic Hierarchy for Information Retrieval, Workshop on English-Chinese Cross Language Information Retrieval, International Conference on Chinese Language Computing, Chicago, IL, pp.187-194.

[5] Chen Keh-Jiann, Shu-Ling Huang, Yueh-Yin Shih, Yi-Jun Chen, 2005, Extended-HowNet- A Representational Framework for Concepts, OntoLex 2005 - Ontologies and Lexical Resources IJCNLP-05 Workshop, Jeju Island, South Korea

[6] 陳克健、黃淑齡、施悅音、陳怡君，2005，多層次概念定義與複雜關係表達－繁體字知網的新增架構，漢語詞彙語義研究的現狀與發展趨勢國際學術研討會，北京大學

[7] 陳怡君、黃淑齡、施悅音、陳克健，2005，繁體字知網架構下之功能詞表達初探，第六屆漢語詞彙語意學研討會，廈門大學

[8] Shu-Ling Huang, Shih Yueh-Yin, Keh-Jiann Chen, 2006, The Knowledge Representation for Comparison Words in Extended-HowNet, CLSW7, NCTU, Taiwan

[9] Yueh-Yin Shih, Shu-Ling Huang, and Keh-Jiann Chen , 2006, Semantic Representation and Composition for Unknown Compounds in E-HowNet, Paclic 20, Wuhan China.

[10] Chia-Hung Tai, Shu-Ling, Huang and Keh-Jiann Chen, 2008, A semantic composition method for Determinative-Measure Compounds in E-HowNet, To appear.