A Step toward Compositional Semantics: E-HowNet a Lexical Semantic Representation System[*]

Keh-Jiann Chen[a] and Shu-Ling Huang[b]

[a]CKIP, Institute of Information Science, Academia Sinica
kchen@iis.sinica.edu.tw
[b]Department of Chinese Language and Literature Studies, National Hsinchu University of Education
slhuang@mail.nhcue.edu.tw

Abstract. The purpose of designing the lexical semantic representation model E-HowNet is for natural language understanding. E-HowNet is a frame-based entity-relation model extended from HowNet to define lexical senses and achieving compositional semantics. The followings are major extension features of E-HowNet to achieve the goal. a) Word senses (concepts) are defined by either primitives or any well-defined concepts and conceptual relations; b) A uniform sense representation model for content words, function words and phrases; c) Semantic relations are explicitly expressed; and d) Near-canonical representations for lexical senses and phrasal senses. We demonstrate the above features and show how coarse-grained semantic composition can be carried out under the framework of E-HowNet. Possible applications of E-HowNet are also suggested. We hope that the ultimate goal of natural language understanding will be accomplished after future improvement and evolution of the current E-HowNet.

Keywords: E-HowNet, compositional semantics, lexical semantic representations

1 Introduction

E-HowNet is a lexical semantic representation model extended from HowNet and hence inherits the basic framework of HowNet. HowNet is an on-line common-sense knowledge base indexing relations of concepts gathered from lexicons of Chinese and English (Dong and Dong, 2006). Each concept is represented and understood by its definition and associated links with other concepts. For instance, the HowNet definition of Warrior|戰士 is as:

{human|人:belong={army|軍隊}, {fight|爭鬥: agent={~}, domain={military|軍}}}

The representation says that a warrior is a human in army who plays the role of agent in the event of military fighting. In HowNet, vocabulary of word sense definition is restricted to a set of around two thousands primitive concepts, called sememes. A word sense is defined by its hypernymy sememe and additional semantic features. HowNet has the following advantages over WordNet (Fellbaum, 1998): 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 for NLP applications. In recent years, HowNet has been applied to a variety of research topics including: word

[*] 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-001-023-MY3.

similarity (劉群, 2002), machine translation (Dong, 1999), and information retrieval (Dorr et al., 2000), etc.

However, what interests us here is how to use HowNet to achieve mechanical natural language understanding. Computer understanding requires a representational framework which represents knowledge about lexical concepts and performs semantic composition and sense disambiguation. So far there has been little research on applying HowNet to semantic composition. We therefore propose a framework extending HowNet, called E-HowNet, to deal with this problem. In the following sections, this paper proposes a variety of inter-connected extensions to HowNet, resulting in E-HowNet. Section 2 describes multi-level concept definitions which enrich its ontological relations. Section 3 discusses methods to achieve semantic composition with a uniform representational framework for both function words and content words. Section 4 defines a precise meaning for *functions* (distinct from the loose use of "function words") as special kinds of relations which maps concepts to concepts. Section 5 addresses some applications of E-HowNet. A summary and conclusions are given in Section 6.

2 Building ontological relations between concepts with multi-level definitions

HowNet uses primitive concepts, called sememes, to define all concepts. For example, '狗 *dog*' is defined as def: {livestock|牲畜}. Using primitives to define concepts causes information degradation and misses important ontological relations between concepts. For example, HowNet defines '獅子狗 *Beijing dog*' as def: {livestock|牲畜} as well. For one, the definition fails to indicate that this species has some connection to Northern China. Secondly the hyponymy relation with 'dog' is lost in their equivalent definitions.

While we've adopted an entity-relational model, i.e. all conceptual relations are explicitly expressed, to define word and phrasal senses, concepts are further defined by both synonyms and simpler concepts, instead of just semantic primitives. Furthermore, all attribute relations are explicitly expressed (陳克健 et al., 2005). In E-HowNet '獅子狗 *Beijing dog*' is defined as def:{dog|狗:source={Beijing|北京}}. This kind of definition explicitly indicates the ontological hyponymy relation between 'dog' and 'Beijing dog' by using the concept 'dog' as the head sense. Thus, the concept definition itself forms an ontological network.

The set of HowNet sememes are also adopted in E-HowNet for ground-level definitions. Since new concepts can be defined by any well-defined concepts, definitions can be dynamically decomposed into lower level representations until the ground-level definition is reached. At that point, all features in the definition are primitive concepts, i.e. sememes. For instance, the top level definition of '文學系 *department of literature*' is shown in (1).

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

The symbol '~', as in HowNet, refers to the head concept of the definition which is 'school department|學系' as in (1). 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').

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

Such a multi-level representational framework makes sense definitions more precise. It also retains the advantage of using semantic primitives to achieve canonical sense representation. Thus, the system aims to achieve both unambiguous definitions and language independence. Followings are additional advantages of multilevel representations.

a. Succinctness and consistency of data management can be maintained through multilevel inheritance of multilevel conceptual hierarchies.

b. The similarity and/or associations of two concepts can be derived by comparing their multilevel representations. Since each level of expansion results in information loss, a multilevel back off process can balance between generality and precision.

c. Near-canonical representation can be achieved at a suitable level of representation for synonyms or paraphrases.

d. New senses are easy to define while maintaining consistency among representations.

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

One of HowNet's objectives is to define word senses that can express synonyms with identical representations. In addition, the similarity of word senses can be derived by comparing definitions. However the E-HowNet's objectives not only intend to achieve the above goals but also to apply semantic composition process to derive the senses representations for phrases and sentences.

HowNet works well for defining content words, but it does not express the senses of function words well. In fact, function words are all defined by the same semantic head: {FuncWord|功能詞:…}. To perform semantic composition, it is essential to have a uniform framework capable of expressing the diversity of both content and function words.

Linguistic theories disagree about the status and or definitions of "function words", but in a general sense, function words are words which typically satisfy syntactic requirements, and may have less referential significance. In fact, it is probably not possible to divide senses, much less words, into independent referential and syntactic uses. However, we can still say that function words indicate 'linguistically' significant relations. For example, the passive '被 *by*' is a preposition that introduces an agent role/relation. Thus, it indicates relations between constituents in a sentence. As well, the morpheme '-ly', in a word like 'gently', creates a bridge between the 'manner' relation of its content sense 'gentle' and whatever verb it connects to. In contrast, content words, particularly nouns, have more independent referentiality and less (or under specified) relational sense. Verbs denote events while also contributing the sense of their argument relations. Nouns refer to objects and also receive roles from predicates.

Thus, anticipating the next section somewhat, E-HowNet treats both nouns and verbs as entities joined by relations of semantic roles. For instance, "John runs"=agent(run, John) which in E-HowNet expression is {run|跑:agent={John}}. This kind of logic is basically equivalent to contemporary event-oriented logic which would write exists:e(run(e) and agent(e,John)).

Thus, we claim that all word interpretations involve two types of sense: relation sense and content sense. However, different syntactic categories could well be said to have different degrees of these senses. A spectrum could be diagrammed as in Table 1. However, for a lexical knowledge representation system to incorporate compositionality, 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. Their sense representations are elaborated in the following section.

Table 1: The sense spectrum for syntactic categories

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

3.1 Lexical representations and basic semantic composition processes

In E-HowNet, the senses of function words are represented by semantic roles/relations (陳怡君 et al., 2005). For example, the conjunction ‘because’ is defined as shown in (2). Function words link two entities, as indicated by x and y below.

(2) *because* 因為

def: reason={}; which means reason(x)={y} (which reads as ‘reason of x is y’) where x is the dependency head and y is the dependency daughter of ‘*because* 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 process:

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 the semantic relation between *A* and *B* to derive relation(*A*)={*B*}.

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

Since methods for word sense disambiguation and relation identification are out of the scope of this paper, we set those issues aside.

Sentence (3) can serve as an example to show how lexical concepts can be combined into a sense representation.

(3) *Because it was raining , the clothes are all wet.* 因為下雨，衣服都濕了。

In the Chinese for the sentence, ‘濕 *wet*’, ‘衣服 *clothes*’ and ‘下雨 *rain*’ are content words while ‘都 *all*’, ‘了 *Le*’ and ‘因為 *because*’ are function words. Their E-HowNet sense representations are shown in Table 2. The main difference in their representations is that the function words are all relations of the form rel(x)=(y). The content words, on the other hand, don’t yet contain information specifying their relations. When a content word acts as a dependency daughter of a head concept, the relation between the head concept and this content word needs to be established by a parsing process. Suppose that the following dependency structure and semantic relations are derived from parsing the sentence (3):

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

(5) is the semantic composition which results from the unification process. The dependency daughters have become feature attributes of the sentential head ‘wet|濕’.

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

In (4), the function word ‘因為 *because*’ links the head concept ‘濕 *wet*’ and ‘下雨 *rain*’ with the ‘reason’ relation. The result of composition is expressed as reason({wet|濕})={rain|下雨}. Other feature representations of the dependent daughters are also inserted according to the step 3 of semantic composition process.

Table 2: Examples of E-HowNet lexical representations

Word	POS	Definition
因為	Cb (conjunction)	reason ={ }
下雨	VA (intransitive verb)	{rain|下雨}
衣服	Na (common noun)	{clothing|衣物}
都	Da (adverb)	Quantity={complete|整}
濕	VH (state verb)	{wet|濕}
了	Ta (particle)	aspect={Vachieve|達成}

3.2 Advantages of uniform representation

The major reason for a uniform representation for both function words and content words is making the process of semantic composition simple and elegant. Other advantages are:

a. Meaning representations for morphemes, words, phrases, and sentences can be uniformly represented under the same framework.

b. Interrogative sentences have representations which, except for the missing information that is queried, are identical to related assertions. Therefore question answering at the conceptual level can be achieved by E-HowNet representations. Take (6) as an example which shows two sentences have very similar E-HowNet representation except that the values of the relation *location* are different. One is asking a place and another provides the answer ‘home|家’.

(6) *Where is he?* 他在何處？
Def:{situated|處於:existent={he|他},location={Ques({place|地方}) }}

He is at home. 他在家。
Def:{situated|處於:existent={he|他},location={home|家}}

c. The sense representations of two expressions also show information similarities and differences with respect to both entities and relations as exemplified in (6).

4 Semantic roles and functions

E-HowNet is an entity-relation model as described above, in which entities indicate objects or events which have content sense, and relations are semantic links between entities. There are two different types of relations: semantic roles and functions (not to be conflated with the general sense of 'function word' described above). 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 'the rel of x is y'. For example, Agent(eat)={the dog} reads 'the agent of the event eating is the dog'. In E-HowNet the sentence of 'the dog eat' is expressed as {eat: agent={the dog}} where 'agent={the dog}' is an abbreviation of 'agent(~)={the dog}'. The symbol '~' denotes the head concept which is 'eat' in this example. A relation rel(x)={y} is considered 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.

Functions are special kinds of relations which map concepts to new concepts in the same domain. This term should not be confused with its more general use in compositional semantics or logic. They do not establish thematic relations or property attributes of concepts, but rather transform concepts into new concepts in the same domain. Functions have compositional properties. New functions can be constructed by composition of many functions of the same type. For instance, the kinship function Father(Father(x)) denotes the grandfather of x and the direction function North(East()) denotes the direction of north-east. Both are composed functions of basic functions. Function expression is written as rel(x). (7) is a typical example.

(7) *vehicle headlight* 車燈

def: {PartOf({LandVehicle|車}): purpose={illuminate|照射: target={road|路}, instrument={~}}}.

In the above definition, 'PartOf' is a function while 'purpose' and 'instrument' are semantic roles. 'Purpose' provides a relation between an event of purpose and its head, in this case by mean of an 'instrument'.

E-HowNet also regards and-or relations, as well as question and negation relations as functions. Their usage parallels familial and direction relations as follows:

(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})}.

In order to achieve the process of automatic feature unification, E-HowNet organizes relations in a hierarchical structure which taxonomically relates entities. A hyponym relation entails its hypernym relations. Taxonomies of semantic roles and sememes are shown in http://ckip.iis.sinica.edu.tw/taxonomy/. The approach advocated here also makes it easy to adopt knowledge represented by other frameworks, such as FrameNet, HowNet etc. Furthermore, concepts can be optionally represented at varying degrees of specificity. Another advantage is that conceptual similarities can be modeled by relational distances in the hierarchy (Resnik, 1999).

5 Applications of E-HowNet

There is still a long way ahead to achieve fully automatic semantic composition and natural language understanding. From syntactic parsing, to semantic composition until canonical sense representation, there are many research problems and difficulties needed to be resolved, such as robust parsing technology, semantic role assignment, sense disambiguation, unknown word identification, aspectual normalization,…, etc. still being hot research topics. E-HowNet didn't solve the problems directly, but it can help in solving those problems. Other than general functions of semantic generalization and specialization, some specific applications of E-HowNet are exemplified below.

In addition to part-of-speeches, conceptual features (sememes in E-HowNet) became prominent features while applying statistical parsing models in resolving ambiguous syntactic structures. We can use conceptual association strength instead of word association strength to overcome data sparseness problem. For instance, even coarse-grained semantic types of event, object, attribute, and attribute-value have significantly different syntactic behaviors. For instance, syntactic ambiguity of *Transitive-Verb + N → Np* or *Vp* is hardly resolved by syntactic rules.

(11) Vp: 檢驗 行李、食物、藥物 v.s. Np: 檢驗 人員、方法、儀器

In (11), 行李 luggage, 食物 food, 藥物 medicine are object type and 人員 agent, 方法 method, 儀器 instrument are attribute type. The above examples indicate the preferences of *Transitive-Verb + Object → Vp and Transitive-Verb + attribute → Np.*

For the task of semantic role assignment, word sense representations of E-HowNet provide ample of conceptual relations as training data for applying machine learning models to determine semantic relation between two dependent concepts/words.

For the task of sense disambiguation, lexical content of E-Hownet provide not only different senses of each word but also preference of conceptual relations, i.e. association strength between two concepts. The above information is essential for the task of sense disambiguation.

In real text, there are 3-5% of unknown words. It is one of major obstacle for natural language understanding. However in Chinese many unknown words are newly coined compounds and they preserve certain degree of semantic compositionality, i.e. their sense and pos can be predicted by semantic composition of morphemes (Chung and Chen, 2009). We have implemented an automatic E-HowNet sense derivation system for determinative-measure compounds by semantic composition method and obtained a very promising result (Tai et al., 2009).

6 Conclusion and future work

HowNet proposed a new model to represent lexical knowledge and inspired us to expand this framework to achieve the task of mechanical natural language understanding. E-HowNet confines each concept to a semantic type, and defines the relation between these types. Hence we have a consistent method for representing concepts, and computers can store and relate meanings.

Semantic composition is a crucial component of linguistic analysis. We have proposed a uniform representation system for both function words and content words to achieve semantic composition, such that meaning representations for morphemes, words, phrases, and sentences can be uniformly represented under the same framework. New concepts can be defined by previously known concepts and definitions can be dynamically decomposed into lower level representations until the ground-level definition is reached. Near-canonical representation thus can be achieved at a suitable level of representation for synonyms or paraphrases. We also suggested compositional functions to extend the expression of new concepts and make word and phrase definitions more detailed and accurate.

There are still many obstacles to achieving the goal of automatically extracting knowledge from language. Apart from sense disambiguation and semantic role labeling, discord between syntactic structures and their associated semantic representations is another critical problem. We need to determine rules which map from coarse syntactic structures to fine-grained semantic relations. Gap filling processes, as discussed here, need to be an integral part of the mechanism. Our future research will continue to address these problems.

References

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

Chung, Yu-San and Keh-Jiann Chen. 2009. Analytic Approach for Sense and Pos Prediction for Chinese Compounds, in preparation.

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

Dong, Zhendong. 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.

Dorr, Bonnie J., Gina-Anne Levow and 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.

Fellbaum, Christiane. 1998. *WORDNET-An Electronic Lexical Database*. The MIT Press.

Resnik, Philip. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research.*

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

Tai, Chia-Hung, Jia-zen Fan, Shu-Ling Huang and Keh-Jiann Chen. 2009. Automatic Sense Derivation for Determinative-Measure Compounds under the Framework of E-HowNet. *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 14, #1, pp.19-44.

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.

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.

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.