Extended-HowNet: A Representational Framework for Concepts

Keh-Jiann Chen, Shu-Ling Huang, Yueh-Yin Shih, Yi-Jun Chen
Institute of Information Science
Academia Sinica, Taipei, Taiwan
kchen@iis.sinica.edu.tw, {josieh, yuehyin, chenyijun}@hp.iis.sinica.edu.tw

Abstract

To bridge the gap between natural language and conceptual representations, we propose a universal concept representation mechanism, called Extended-HowNet, which was evolved from HowNet. It extends the word sense definition mechanism of HowNet and uses WordNet synsets as vocabulary to describe concepts. Each word sense (or concept) is defined by some simpler concepts. The simple concepts used in the definitions can be further decomposed into even simpler concepts, until primitive or basic concepts are obtained. In this way, definitions can be dynamically decomposed and unified into Extended-HowNet representations at different levels. Extended-HowNet is language independent; thus, any word sense of any language can be defined and near-canonical representation can be achieved. Given any two concepts, not only their semantic distances, but also their sense similarities and differences can be derived by comparing their definitions. In addition to taxonomy links, concepts are also associated by their shared conceptual features, while fine-grain differences among near-synonyms can be differentiated by adding new features.

1. Introduction

An ontology is a specification of a conceptualization (Gruber, 1993). We propose a frame-based entity-relation knowledge representation model called Extended-HowNet, which was evolved from HowNet (Dong & Dong, http://www.keenage.com/), to encode concepts, such as

<science fiction> Def:=
{<book>: content = {<imagination>:
domain= {<science>}}}.

This says that a science fiction entry is a book with imaginary content in the science domain. The objective of Extended-HowNet is to achieve near canonical conceptual representation and semantic composition capabilities. In Extended-HowNet, concepts are represented and understood by their definitions and associated links to other concepts. We define each lexical sense by well-defined concepts, which are not necessarily primitive concepts. The vocabularies used for definitions can be replaced by WordNet synsets (Fellbaum, 1998). The advantage of using WordNet synsets is that they help achieve universal and language independent representations, since each synset has a unique sense and the sense similarity between two synsets can be measured through WordNet's ontology.

The remainder of this paper is organized as follows. In Section 2, we introduce related works on lexical knowledge representation. Section 3 gives the formal definition of Extended-HowNet, after which the advantages of the system are described in Section 4. Finally, we summarize our work and present our conclusions in Section 5.

2. Background

To achieve natural language understanding, computer systems need to know the sense similarity and dissimilarity of two sentences or two words. This requires the support of ontologies that:

a) Identify synonym concepts and measure the similarity distance between two concepts.
b) Know the shared semantic features and feature differences between two concepts.
c) Provide a unique index of each concept, such that associated knowledge can be coded and accessed.
d) Utilize language independent sense encoding.
e) Make logical inferences through a conceptual property inheritance system.
f) Incorporate dynamic concept decomposition and composition mechanisms.

None of the current ontologies provide all the above functions. We therefore propose a sense

representation framework extended from HowNet that meets this need.

2.1 WordNet-like ontologies

WordNet (Fellbaum, 1998) contains information about nouns, verbs, adjectives, and adverbs in English and is organized around the notion of a synset. A synset, which roughly denotes a concept, is a set of words with the same parts-of-speech that can be interchanged in certain contexts. For example, {car, auto, automobile, and motorcar} form a synset, because they can be used to refer to the same concept. Synsets can be related to each other by semantic relations, such as hyponymy, meronymy, or cause. Furthermore, a synset is often described by a gloss, such as "4-wheeled; usually propelled by an internal combustion engine".

2.2 HowNet

HowNet is an on-line common-sense knowledge base that provides the inter-conceptual relations and inter-attribute relations of concepts found in lexicons of the Chinese language and their English equivalents. An introduction to HowNet can be found at http://www.keenage.com/zhiwang/e_zhiwang.html. In HowNet, word sense representations, also called definitions, are encoded by a set of approximately two thousand primitive concepts, called sememes. A word sense is defined by its hypernymy sememe and additional semantic features. For instance, the HowNet definition of Warrior|戰士 is:

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

which says that a warrior is a human in an army who plays the role of an agent in the event of military fighting.

3. Extended-HowNet

Using primitives to encode the meaning of a concept causes the information to be degraded so that it is almost impossible to understand the representation of a complex concept. Furthermore, it is debatable whether there exists a limited and fixed set of so-called primitives. In Extended-HowNet, we adopt a similar mechanism as HowNet to define word senses except that a concept is defined by simpler or synonym concepts, instead of semantic primitives only. Thus, <man> is a <human> of <male> gender is defined as <man>:={<human>: gender={<male>}} in our extended system. The sememes used in the current version of HowNet are also adopted as ground-level definitions in Extended-HowNet. In our proposed system, new concepts are defined by any well-defined concepts, and a definition can be dynamically decomposed into lower level representations to find the ground-level definition in which all the features are sememes. For instance, the top level definition of <department of literature|文學系> is {<school department|學系>: predication= {<teach| 教 >: location={~}, theme={<literature| 文 >}}}. Since the concept <school department| 學 系 > is not a primitive concept, the above definition can be further extended to the ground-level definition, {<InstitutePlace|場所>: domain = {<education|教育 >}, predication= {<study| 學 習 >: location= {~}}, predication= {<teach|教>: location = {~}, theme={<literature|文>}}}. In order to describe precise definitions of concepts, a number of technical problems must be solved. First, to achieve unambiguous definitions, each referred concept should be unambiguous. For this reason, we adopt WordNet synsets as the vocabulary for conceptual indexing and representation in Extended-HowNet. Second, it is necessary to determine which major features of a concept are sufficient to define the concept. We discuss this issue in detail in Section 3.1. Third, semantic composition and decomposition involve feature unification, so feature values under the same relation type should be combined during the unification process, Thus, in the above example, the hypernym class <school department> of <department of literature| 文 學 系 > is not a primitive concept and can be extended to the definition of {<InstitutePlace| 場 所 >: domain= {<education|教育>}, predication={<study|學習>: location={~}}, predication= {<teach| 教 >: location={~}}. The reduplicated features of predication={<teach| 教 >:location={~}, theme= {<literature|文>}} can then be combined.

Formally, Extended-HowNet is a quadruple feature unification system comprised of Vocabulary, Grammar, Taxonomy of Concepts, and Taxonomy of Relations, where:

a) Vocabulary= WordNet Synsets;
b) Grammar= the syntax for Extended-HowNet;
c) Taxonomy of Concepts= the hierarchical structure of concepts and sememes formed by hyponym relations and part-whole relations; and
d) Taxonomy of Relations= the hierarchical structure of the relations formed by hyponym relations.

In Extended-HowNet, our goal is to unify the WordNet and HowNet taxonomies into a taxonomy of concepts; and combine the semantic relations of FrameNet and HowNet to form a taxonomy of relations.

3.1 Principles of Concept Definition

The meaning of a concept is supported by its associated concepts, including its formal properties, constituents, purposes, and relations to other concepts. When defining a concept, it impossible to encode all of its associated relations. Thus, ,the principle for defining a concept is to first identify its immediate hypernym and then encode its most important features that suffice to differentiate the concept from other concepts. The qualia structure is the major feature for a nominal-type concept (Pustejovsky, 1995) and an event frame is the major feature of an event-type concept (Fillmore, *FrameNet*). The qualia of an object, defined by Pustejovsky (1995) are:

a) Constitutive: the relations between the object and its constituents, such as its materials, parts, and components.
b) Formal: the properties to distinguish the object within a larger domain, such as its shape, magnitude, and color.
c) Telic: the purpose and function of the object.
d) Agentive: the factors involved in the origin or "bringing about" of the object.

There are two types of attribute feature: 1) a simplex attribute, which is a feature-value pair expressed by some sememes; and 2) a complex relative clause, which is an event frame comprised of eventive features. The constitutive and formal properties in Extended-HowNet are represented by simple attribute-value pairs, i.e., Relation={Concept} pairs, while the telic and agentive properties are usually represented by event frames. For example, the concepts of <teacher> and <student> may be defined and differentiated as <teacher>:= {<human>: telic={<teach>:agent={~}}} and <student>:= {<human>: telic= {<teach>: goal={~}}}. An event-type concept is also defined by its hypernym event-type. Brotherhood concepts are differentiated by their event frame elements, including participant roles and adjuncts, as well as their semantic restrictions.

3.2 Consistency and Integrity of Representations

The integrity of concept representation is supported by the dynamic conceptual associations within the Extended-HowNet system. As we know, the meaning of a word is expressed by its associations with other concepts. Therefore, in Extended-HowNet, a concept is associated to other concepts through

a) Taxonomies, such as SUMO (Niles & Pease, 2001), WordNet, HowNet, FrameNet, SIMPLE-CLIPS, and EuroWordNet. The association relations include synonymy, hyponymy, antonymy, and meronymy.

b) Dynamic definition extensions: High-level features (i.e., concepts) provide easy encoding for general knowledge. Usually, important conceptual properties are associated with basic concepts, not primitive concepts. For instance, Pluto is a dog, and "dog" is a basic concept. Although it is possible to define a basic concept by primitive concepts, it does not really help us understand the basic concept. For example, the associated properties of dogs, such as "dogs bark", "dogs are pets"… are hardly associated with the primitive concept of 'animal'.

3.3 Feature Inheritance and Conceptual Extension

The meaning of a concept is supported by its associated concepts. Clearly, the associated properties or knowledge of a particular concept can be accessed or encoded directly through its definition, or indirectly inherited from its ancestors. Furthermore, a hierarchical taxonomy also provides a semantic distance between two

concepts. However, conventional taxonomies do not provide the exact semantic similarities and dissimilarities of two concepts. In contrast, the definitions of concepts in Extended-HowNet not only provide taxonomy and semantic similarities, but also encode the semantic differences between two concepts.

Taxonomically unrelated but conceptually related concepts can also be computably associated through Extended-HowNet. The graphical relations in Figure 1, reproduced from HowNet (Dong & Dong, http://www.keenage.com/), show the concepts that may not be associated with taxonomical relations, but may be associated with each other by other semantic relations.

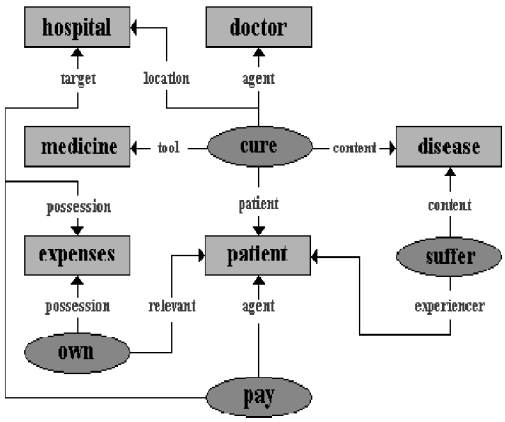

Figure 1. Concepts associated by their semantic relations

3.4 Difficulties and Solutions

Some types of concept do not have natural hypernym concepts, as is usually the case with the parts of an object. Instead, they are linked to other concepts by part-whole relations in the ontology. We expand Extended-HowNet with a new notation, %, to denote *part-of*, and use the feature relations of the location or telic to distinguish different parts. For example:

<foot|腳> Def:=
{ %<animal|獸>:
telic={<walk|行走> :
agent = {~}}.

The definitions of relational-type concepts, such as kinship relations and directional relations, are different from the definitions of entities. For instance, <grandfather> and <north-west> have to be expressed by composing primitive relations, instead of feature attributes.

<grandfather>Def:= {father(father(human:x))}
<north-west> Def:={north(west(location:x))}

A detailed discussion of this point can be found in Chen et al. (2004).

Functional-type concepts, such as adverbs, prepositions, and conjunctions, contain fewer content senses, but rich relational senses. Definitions of function words cannot be based solely on their parts-of-speech, since the latter do not provide semantic information and cannot fit into the unification process for semantic composition. Function words are defined by their relational senses and content senses (Chen et al., 2005). For example, the adverb <in public|當眾> is defined as

Def:= manner={overt|公開},

and the preposition <by|被> is defined as

Def:= agent={}.

It is necessary to distinguish between individual instances and generic concepts, as proper names refer to individual not generic concepts. We therefore use the notation (<concept>), instead of {<concept>}, to denote an individual instance of <concept>. For example,

<Tomas Edison|愛迪生 > Def:=
(< scientist| 科學家>:
name= 'Tomas Edison|愛迪生', …)

Some concepts are hard to define by common concepts. For instance, concepts like <square root>, <prime number>, <gravity>, and <palm tree> that belong to certain special domains are hard to define in detail, because they require the support of a domain ontology and domain knowledge. Currently, we do not provide detailed definitions of domain specific concepts in Extended-HowNet; however we will try to link them to a domain ontology in our future work.

4. Advantages of Extended-HowNet

The following advantages of Extended-HowNet show how it bridges the gap between string processing and conceptual processing.

a) Feature representation is more precise and incremental. e.g.,

<great dane|大丹狗>Def:=
{<dog|狗>:

place={<German|德國>},
telic={<hunt|狩獵>:
instrument={~},
size={<big|大型>},
evaluation={<gentle|溫和> },
color={<black white|黑白>}}}.

Note that a pure taxonomy approach, such as WordNet, does not provide a detailed description of a concept.

b) Features are used as the criteria for classifying new types. For example, <great dane> is also classified as：

1) A hunting instrument (according to its telicity feature); another example of the class is

<firearm> Def:=
{<gun|槍>:
telic={<hunt|狩獵>:
instrument={~}}}

2) Animals with black/white colors; another example of the class is

<zebra> Def:=
{<horse|馬>:
color={<black white|黑白>}}.

c) The system can achieve near canonical semantic representation. Thus, two sentences with different surface forms or in different languages may have similar Extended-HowNet representations. e.g.,

a) 我 買了 一本 科幻小說。
b) I bought a science fiction book.

Two sentences in different languages can have the same representation of {<buy|買>: agent={<I|我>}, goal={<science fiction|科幻小說>: quantity={<one|一>}, time-before = {speaking time}}. Note that the above high-level representation can be extended to lower level and WordNet synset representations.

d) Extended-HowNet enables multi-level meaning decomposition. e.g.,

<tailor store|裁縫店> Def:=
{<store|店>:
telic={<sew|裁縫>:
location={~}}},

which can be extended to

{<InstitutePlace|場所>:
{<produce|製造>:
PatientProduct=
{<clothing|衣物>},
location={~}}}.

In contrast, HowNet concepts are defined by primitive concepts. Thus, in the above example, the basic concept <InstitutePlace|場所> does not include the information about "commerce" in <store|店>.

e) Extended-HowNet is universal and language independent, since it uses WordNet synsets as its descriptive language.

f) Extended-HowNet does not create a completely new ontology, but accommodates other ontologies, such as WordNet, HowNet, and FrameNet, instead.

5. Summary and Conclusion

To bridge the gap between natural language representations and conceptual representations, we have proposed a universal concept representational mechanism, called Extended-HowNet, which uses the word sense definition mechanism of HowNet and WordNet's synsets as vocabulary to describe concepts. Fine-grain differences among near-synonyms can be differentiated by adding new features. The encoded features, including qualia structures and ontological links, provide the bases for manipulating intelligent semantic processes, such as type coercion, semantic composition, rule generalization, and logical inference. The semantic distance of two concepts can be computed by their Extended-HowNet representational distance.

In addition to conventional taxonomic relational links, such as synonymy, hyponymy, antonymy, and meronymy, Extended-HowNet also links concepts by their shared features. Multiple links mean multiple-inheritances. In Extended-HowNet, the shared properties of different concepts are associated with common ancestors without redundancy.

Extended-HowNet is language independent; thus, it can bridge the gap in translation equivalence between two languages. In EuroWordNet, each word sense of a different language is intended to link to a synonymous WordNet synset. However, as many word senses cannot find such a synset, they have to create some interlingua-indices (ILI) to link translation equivalences among different languages (Vossen, 2000). If WordNet synsets and the senses of all

ILI were defined in Extended-HowNet, it would become a shared ontology for all languages.

As Extended-HowNet is universal, it can define any concept. However, there are still some problems that we must address. For example, concepts (such as 'square root', 'prime number', 'gravity', and 'palm tree') that belong to certain special domains are complicated and hard to define. In addition to their definitions, the representation of such concepts needs support from related domain knowledge-bases.

The semantic composition and decomposition mechanism in Extended-HowNet can be extended to encode the deep semantics of phrases and sentences. Detailed representations of references, quantifications, and temporal relations will be addressed in our future work. Finally, motivated by the syntactic differences within synonyms (Levin, 1993; Huang et al., 2000; Chen et al., 2005), we will incorporate fine-grain features, in particular semantic/syntactic correlation features, into future refinements of Extended-HowNet.

Acknowledgements: This research was supported in part by National Science Council under a Center Excellence Grant NSC 93-2752-E-001-001-PAE and Grant NSC93-2213-E-001-019. We especially thank Prof. Dong Zhengdong for developing HowNet and providing its lexical database.

References

Chen, Keh-Jiann, Shu-Ling Huang, Yueh-Yin Shih, Yi-Jun Chen, 2004, *Multi-level Definitions and Complex Relations in Extended-HowNet*, Workshop on Chinese Lexical Semantics, Beijing University. (in Chinese)

Chen, Yi-Jun, Shu-Ling Huang, Yueh-Yin Shih, Keh-Jiann Chen, 2005, *Semantic Representation and Definitions for Function Words in Extended-HowNet*, Workshop on Chinese Lexical Semantics, Xiamen University. (in Chinese)

Dong, Zhendong & Dong Qiang, *HowNet*, http://www.keenage.com/

Dowty, David R. 1991, "Semantic Proto-roles and Argument Selection", *Language*, Vol. 67(3), pp. 547-619.

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

Fillmore, Charles, *FrameNet*, http://www.icsi.berkeley.edu/~framenet/

Gruber, T.R. 1993, *Toward principles for the design of ontologies used for knowledge sharing*, Padua workshop on Formal Ontology.

Huang, Chu-ren, K. Ahrens, Li-li Chang, Keh-Jiann Chen, M. C. Liu, Mei-Chih Tsai, 2000, " The Module-Attribute Representation of Verbal semantics: From Semantics to Argument Structure" *International Journal of Computational Linguistics and Chinese Language Processing*, Vol.5, #1, pp.19-46.

Levin, Beth. 1993, *English Verb Classes and Alternations: a Preliminary Investigation*, Chicago Press.

Niles, Ian and Pease, Adam. 2001."Towards a Standard Upper Ontology," *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems* (FOIS-2001), Ogunquit, Maine, October 17-19.Pustejovsky, James 1995, *The Generative Lexicon*, the MIT press.

Schank, R., 1975, *Conceptual Information Processing*, Amsterdam, North-Holland.

Sowa, John, 2000, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co..

Vossen, Piek (ed.), 2000, *EuroWordNet General Document*, http://www.hum.uva.nl/~ewn.

Wiezbicka, A. 1972, *Semantic Primitives*, Athenaum, Frankfurt.