E-HowNet and Automatic Construction of a Lexical Ontology

Wei-Te Chen, Su-Chu Lin, Shu-Ling Huang, You-Shan Chung, and **Keh-Jiann Chen**
Institute of Information Science, Academia Sinica
weitehchen@gmail.com,
{jess, yosieh, yschung, kchen}@iis.sinica.edu.tw

Abstract

In this paper, we propose a lexical senses representation system called E-HowNet, in which the lexical senses are defined by basic concepts. As a result, the meanings of expressions are more specific than those derived by using primitives. We also design an ontology to express the taxonomic relations between concepts and the attributes of concepts. To establish the taxonomic relations between word senses, we introduce a strategy that constructs the E-HowNet ontology automatically. We then implement the lexical ontology as a Web application[1] to demonstrate the taxonomy and the search functions for querying key-terms and E-HowNet expressions in the lexicon, which contains more than 88,000 lexical senses.

1 Introduction

E-HowNet, an evolution and extension of HowNet (Dong & Dong, 2006), is an entity-relation representation model for lexical senses. Under the framework, word senses are defined by basic concepts as well as conceptual relations called attribute-values. The following is an example of lexical sense representation in E-HowNet.

(1) '慎 選|carefully choose' is expressed (or defined) by the expression '{choose|選擇:manner={cautious|慎}}'.

In the representation, the meaning of "慎選" is comprised of two primitive concepts, "choose|選擇" and "cautious|慎", and the conceptual relation between the primitives is explained by the semantic role "manner". For further details, readers may refer to the E-HowNet technical report (CKIP 2009).

With a well-established entity-relation model, semantic composition is applicable from the morphological level to the sentential level in E-HowNet. Semantic compositionality, together with syntactic information, contributes enormously to natural language understanding.

The remainder of this paper is organized as follows. We describe the major features of E-HowNet in Section 2 and introduce the E-HowNet ontology in Section 3. Then, we present our online E-HowNet system in Section 4. Section 5 contains some concluding remarks.

To achieve the goal of semantic compositionality and to extend the advantage from HowNet, the following features are implemented in E-HowNet.

a) Multi-level definitions and semantic decomposition: Word senses (concepts) can be defined (expressed) by primitives as well as by any well-defined concepts and conceptual relations. However, using only primitives to express word senses, as in HowNet, causes information degradation and important ontological relations between concepts may be missed.

b) Uniform sense representation and semantic compositionality: To achieve semantic compositionality, it is necessary to encode the senses of both content words and function words in a uniform framework. HowNet performs well for defining content words, but it does not provide a well-form representational framework for expression the sense of function words, which indicate semantic relations. In contrast, E-HowNet

[1] available at http://ckip.iis.sinica.edu.tw/~wtchen/taxonomy/

Coling 2010: Demonstration Volume, pages 45–48,
Beijing, August 2010

provides uniform representations for the senses of content/function words and the senses of sentences/phrases. For example, the passive sense of the preposition '被 by' introduces an agent role (relation) and the conjunction '因為 because' links the relation of reason between two events. The functional representation and semantic compositionality are illustrated by the following example:

(2) Because of the rain, the clothes are all wet.
因為下雨，衣服都濕了。

Table 1: The function representation and semantic compositionality for example sentence

Word	POS	E-HowNet 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}

Suppose that the following dependency structure and semantic relations are derived by parsing sentence (2) as follows:

(3) 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)。

The semantic composition in (4) is the result of unifying the features of the lexical representations shown in the above table. The dependency daughters have become feature attributes of the sentential head 'wet|濕'.

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

c) Taxonomy for both entities and relations: To achieve automatic feature unification, E-HowNet organizes entities and relations (attributes) in a hierarchical structure that relates entities taxonomically. Further details are provided in the next section.

2 Ontology

We adopt and extend approximately 2,600 primitives from HowNet to form the top-level ontology of E-HowNet, which includes two types of subtrees: entities and relations. The entities are comprised of events, objects, and attribute-values; while the relations are comprised of semantic-roles and functions. Entities indicate concepts that have substantial content, whereas relations link the semantic relations between entities (Chen et al., 2004; Chen et al., 2005; Chen et al., 2005; Huang et al, 2008). The taxonomic structure is organized by hypernym-hyponym relations; therefore, it forms an inheritable system, i.e., the hyponym concepts inherit the properties of hypernym concepts. The proposed approach facilitates the adoption of knowledge represented by other frameworks, such as FrameNet, and HowNet; and it allows concepts to be represented with varying degrees of specificity. Another advantage is that conceptual similarities can be modeled by their relational distances in the hierarchy (Resnik, 1999), and the taxonomic relations between lexical senses can be captured from their E-HowNet expressions automatically.

2.1 Automatic Construction of Ontology

With E-HowNet expressions, lexical senses are defined as entities and relations. Thus, all the taxonomic relations of lexical senses can be identified according to their E-HowNet definitions. Synonyms are identified by their identical E-HowNet expressions, and hyponymy relations are identified by the subsumption of attribute-values. (Note that only near-synonym classes are identified due to the coarse-grained expressions of the lexical senses in the current version of E-HowNet.) Furthermore, new categories are identified by common attribute-values. For instance, pandas and zebras can be categorized as animals with the same feature: black and white markings. To construct a complete lexical taxonomy, we use

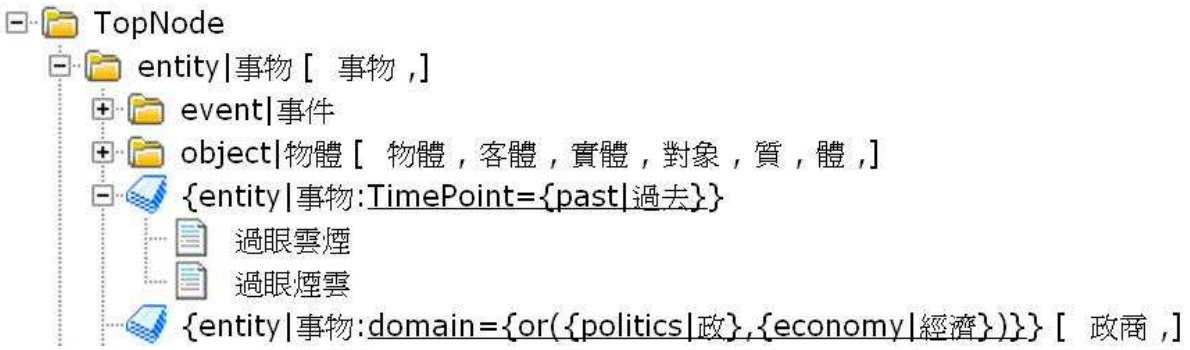

Figure 1: The E-HowNet ontology system

a strategy that categorizes concepts automatically.

Starting with a manually created top-level ontology of primitive concepts, the following strategy classifies the lexicon into hierarchical subcategories:

(1) Attach lexical senses. Words and associated sense expressions are first attached to the top-level ontology nodes according to their head concepts. For instance, the head concept of the expression '{choose|選擇:manner={cautious|慎}}' is 'choose|選擇'.

(2) Sub-categorization by attribute-values. Lexical concepts with the same semantic head are further sub-categorized according to their attribute-values. Lexicons that have the same attribute-values share specific characteristics; therefore further sub-categorization is performed based on the distinct attribute-values of the lexicons.

(3) Repeat step (2) if there are too many lexical concepts in one category. Although the lexicons are classified after step (2), some sub-categories might still contain too many lexicons. In this situation, we further classify the lexicons in the sub-category with other attribute-values until all sub-categories contain fewer members than a predefined threshold, or all members of a category are synonyms.

3 Overview of the On-line System

The current E-HowNet ontology is an on-line version of the automatically constructed taxonomic structure of E-HowNet expressions, which contain more than 88,000 lexical senses. This section provides an overview of the ontology and the functions of the on-line web browsing system.

Figure 2: Key-Term Search Box

Figure 1 shows the E-HowNet ontology system and tree structure.

The tree structure of hyponymy relations allows users to browse the entire tree by expanding and hiding sub-trees. Although the classification strategy enables the number of entities under each node to be limited and viewed easily, a more effective function is essential for exploring more than 88 thousand items of data in E-HowNet. Therefore, we provide a search function that allows users to query lexical senses in two ways:

Key-Term Search: The first way is key-term search, which is shown in Figure 2. The syntax of the query interface is like that used by conventional search engines. By inputting the key-term "物體" , the system will search all the taxonomy nodes, sub-categories, and lexical nodes. Then, the results for the taxonomy node "object|物體" and the lexical word "物體" will be displayed in

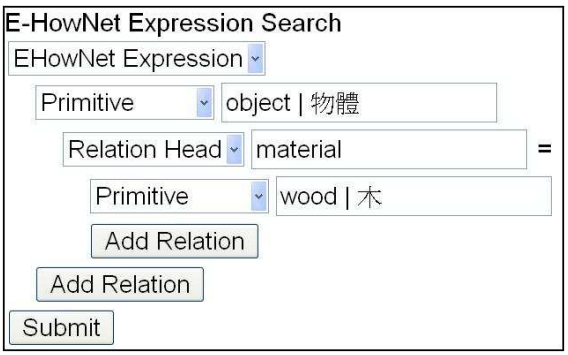

Figure 3: E-HowNet Expression Search Box

the respective columns.

E-HowNet Expression Search: To search a class of words with specific attribute-values, we provide another query syntax for exploring data in E-HowNet Expression. For instance, to find all expressions about wooden objects involves finding E-HowNet data items containing the entity "object—物體" and the attribute-value "material={wood|木}". The expressions are entered on the form shown in Figure 3 and submitted to the system. The results of word senses denoting wooden objects are then returned.

4 Conclusion

E-HowNet sense representations are incremental. Hence, lexical sense expressions can be updated and refined at anytime. In addition, logical relations and the taxonomic structure can be rebuilt automatically based on the refined expressions. New categories in the taxonomy can be identified and characterized by their specific attribute-values. Uniform representations of function words and content words facilitate semantic composition and decomposition, and allow users to derive sense representations of phrases/sentences from the composition of lexical senses. Furthermore, because of E-HowNet's semantic decomposition capability, the primitive representations for surface sentences with the same deep semantics are nearly canonical. We have implemented the E-HowNet ontology online to demonstrate the taxonomy, sub-categories, and lexicons in a hierarchical tree structure. In addition, we provide search functions for querying key-terms and E-HowNet expressions.

References

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

Keh-Jiann Chen, Shu-Ling Huang and Yueh-Yin Shih, Yi-Jun Chen. 2005. Extended-HowNet- A Representational Framework for Concepts. In *Proceedings of OntoLex 2005*, Jeju Island, South Korea.

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

Z. D. Dong and Q. Dong 2006. HowNet and the Computation of Meaning. World Scientific Publishing Co. Pte. Ltd.

Shu-Ling Huang, Shih Yueh-Yin and Keh-Jiann Chen 2008. Knowledge Representation for Comparison Words in Extended-HowNet. *Language and Linguistics*, vol. 9(2), pp. 395-414.

Philip Resnik. 1999. Semantic similarity in a Taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artifical Intelligence Research*, vol. 11, pp. 95-130.

CKIP. 2009. Lexical Semantic Representation and Semantic Composition: An Introduction to E-HowNet (E-HowNet Technical Report). Academia Sinica, Taipei.