Semantic Representation and Composition for Unknown Compounds in E-HowNet

Yueh-Yin Shih, Shu-Ling Huang, and Keh-Jiann Chen

Institute of Information Science,
Academia Sinica, Taipei, Taiwan
{yuehyin,josieh}@hp.iis.sinica.edu.tw, kchen@iis.sinica.edu.tw

Abstract. This paper describes a universal concept representational mechanism called E-HowNet, to handle difficulties caused by unknown words in natural language processing. Semantic structures and sense disambiguation of unknown words are discovered by analogy. We intend to achieve that any concept can be defined by E-HowNet and the representation is near-canonical. The design for easy semantic composition and decomposition makes the automation of semantic processing for unknown words, phrases and even sentences possible.

Keywords: unknown word, compound, E-HowNet, semantic composition, analogy, semantic similarity.

1 Introduction

The occurrences of unknown words cause difficulties in natural language processing. The word set of a natural language is open-ended. Because novel words are created daily for expressing new concepts and new inventions, it is impossible to collect them completely in a dictionary. Morphology of Chinese allows new words to be generated by compounding and affixation. Orthographically, such compounding and affixation are represented by combinations of characters. In general, each Chinese character represents a morpheme. Each Chinese morpheme carries meanings and most are polysemous. New words are easily constructed by combining morphemes and their meanings are the semantic composition of morpheme components. (Of course there are also exceptions of semantically non-compositionally compounds.) It is impractical to construct a lexical database containing semantic information for all words. Therefore, it would be valuable to have a good design of semantic representation system which can easily compose the necessary information for new words.

A compound is a combination of two or more words/morphemes bounded together by morpho-syntactic relations such as modifier-head, verb-object and so on. According to an inspection on the Sinica Corpus [1], 3.51% of the word tokens in the corpus are unknown, which contains about 80,000 entries. Among them, about 51% of the word types are compound nouns, 34% are compound verbs and 15% are proper names. In this paper, we introduce our concept representation mechanism, called E-HowNet and demonstrate the capabilities of this system in representing senses for unknown compound nouns, due to its advantage of easy semantic composition and decomposition.

This paper is organized as follows. Section 2 describes the characteristics of E-HowNet. In section 3, unknown compound nouns are taken as examples to demonstrate the analogy approach in discovering the semantic structures of unknown words in E-HowNet. Conclusion and future work are drawn in section 4.

2 E-HowNet

Lexical knowledge representation has become a major research area for natural language processing in recent years. To bridge gaps between natural language representations and conceptual representations,

we had proposed a frame-based entity-relation knowledge representation model called E-HowNet, which was evolved from HowNet [2] to encode concepts. It extends the word sense definition mechanism of HowNet and uses WordNet synsets [3] as vocabulary to describe concepts. In principle, the qualia structure is the major features for a nominal-type concept [4] and event frames are for eventive concepts [5]. For example, we define concepts "sound recorder" as (1) below:

(1) 錄音機 "sound recorder"
def:{machine|機器:
telic={voice recording|錄音:
instrument={~}}}

Under the mechanism, E-HowNet links concepts by the conventional taxonomic relation links, such as synonymy, hyponymy/hypernymy, antonymy, meronymy and their shared features. In the above example, "machine" is the hypernymy of "sound recorder". The qualia "telic" denotes the purpose and function of "sound recorder" is the event "sound recording". The instrument for sound recording is "~" which denotes the head concept of the definition, i.e. "machine|機器. Thus, the definition can be glossed as: "a sound recorder is a machine which functions as the instrument of sound-recording activity".

Complex concepts are represented by simpler concepts which are not necessary to be primitive concepts. The simple concepts used in the definitions can be further decomposed into even simpler concepts, until primitive or basic concepts are reached. Therefore the representation of a concept can be dynamically decomposed and unified into different levels of representations [6, 7]. In (1), the concept "sound recorder" is defined by simpler concepts "machine" and "sound recording". The concept "sound recording" is not a primitive concept and can be further decomposed into primitive concepts "record" and "sound" as in (2). Thus the definition in (1) can be extended to reach the ground-level representation as in (3). Such multi-level representations are easier to understand, for it accords with human cognition models.

(2) 錄音 "sound recording"
def: {record|記錄:content={sound|聲}}

(3) 錄音機 "sound recorder"
def: {machine|機器:
telic={record|記錄:
content={sound|聲},
instrument={~}}}

In E-HowNet, we intend to unify WordNet and HowNet taxonomies as the taxonomy of concepts and combine the semantic relations of FrameNet and HowNet to form the taxonomy of relations. The processes of semantic decomposition and feature unification rely on these taxonomies.

3 Discovering Semantic Structures by Analogy for Compound Nouns in E-HowNet

Veale[8] tests the ability of HowNet system in doing analogy generation and concludes that HowNet contains sufficient structure to realistically support both a taxonomic abstraction view and a structure-mapping view of analogy generation. Since E-HowNet adopts and extends the sense definition mechanism of HowNet, we will use similar strategy to discover the semantic structures of a very productive type of unknown words, i.e. compound nouns.

E-HowNet uses hypernym concepts as the type classifications for concepts and differentiates concepts of same hypernym class by their major features. To discover the sense and semantic structure of a noun compound is to disambiguate the semantic ambiguity of the morphological head of a

compound noun and find the proper semantic relation between constituents of the compound. For example, when we see the unknown/undefined compounds such as 牧工 “hired herdsman”, 核工 “nuclear industry” ,or 唱工 “art of singing”, firstly, we have to find the appropriate meaning for each head of these unknown compound. Secondly, we have to build the correct relation between their modifiers and the heads, such as the relation between 牧 and 工, 核 and 工, etc.

Chen & Chen [9] proposed an example-based similarity measure to disambiguate the polysemous heads. They extracted some examples with the polysemous head morpheme from corpora and dictionaries, and classified them into different groups according to their meaning. Let’s take “工” as example and add E-HowNet definitions for each class, shown as table 1:

Table 1.

Sense	**example**	**E-HowNet definition**
工人 “labor”	搬運工 “porter”	def:{labor|工人: telic={transport|運送: patient={goods|貨物}, agent={~}}}
	女工 “female labor”	def:{labor|工人:gender={female|女}}
	童工 “child labor”	def:{labor|工人:age={child|幼兒}}
工業 “industry”	化工 “chemical industry”	def:{industry|工業: domain={chemistry|化學}}
	機工 “engineering industry”	def: {industry|工業: domain={machine|機器}}
技術 “skill”	刀工 “cutting skill”	def:{skill|技術: predication={cut|切削: method={~}}}
	畫工 “painting skill”	def:{skill|技術: predication={draw|畫: method={~}}}

The meaning of 牧工 “hired herdsman”, 核工 “nuclear industry” ,or 唱工 “art of singing” are then determined by comparing the similarity between their modifiers and the modifiers of each class of examples. That is, we compare 牧、核 and唱 separately with 搬運、女、童、化、機、刀...etc. And then choose the most similar meaning as their head meaning. For instance, 牧 is most similar with the modifiers in first class, thus the head of 牧工 is “labor”. As for the semantic similarity calculation, several measures of semantic similarity have been proposed for taxonomy based dictionary.[1]

Based on similarity calculation, a preliminary definition can be made for each unknown/undefined compound. To further define them, we need to know the relation between the modifiers and their head. Suppose we know the examples in class two are all defined by the same semantic feature “domain”, then we can further define 核工 as (4):

(4) 核工 “nuclear industry”
def:{industry|工業:domain={nucleonics |核子學}}

[1] Such as Wu & Palmer, *Verb semantics and lexical selection*. (Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics), Leacock & Chodorow, *Combining local context and WordNet similarity for word sense identification*. (WordNet: An Electronic Lexical Database), Resnik, *Semantic similarity in a taxonomy.* (Journal of Artificial Intelligence Research 11), Lin, *An information-theoretic definition of similarity.* (ICML-98) ,Jiang & Conrath, *Semantic similarity based on corpus statistics and lexical taxonomy*. (ROCLING X), Lesk, *Automatic sense disambiguation using machine readable dictionaries.* (Proceedings of the Special Interest Group for Design of Communications Conference).

We only need to replace the value of domain with the modifier of undefined compound “核” to create a new definition in class two. In the same way, 唱工 can be defined as (5):

(5) 唱工 “art of singing”
def:{skill|技術:predication={sing|唱:method={~}}}

But, in general, the relation between modifier and head is also ambiguous. For instance, for the sense class of “labor|工人”, there are three different definition patterns lead by the features of “telic”, “gender”, and “age” in the same class. To decide which pattern is the best one, the same reasoning strategy of comparing the similarity of the modifiers is adopted. We compare 牧 with the defined modifiers搬運 “transport”, 噴漆 “spray paint”, 焊接 “weld”...etc. which are all linked by the relation of “telic” in class one. And then compare 牧with modifier女 “female” and男 “male” which are related with “gender”, and at last, we compare 牧with modifier童 “child”, 少年 “teenage”, 青壯 “adult”, 老年 “aged”...etc. which are all linked in the relation of “age”. The similarity shows that the features linked by the relation “telic” is most similar to 牧. Therefore the correct relation between 牧 and 工 is established which helps to derive the sense representation of 牧工 as (6):

(6) 牧工 “hired herdsman”
def:{labor|工人:telic={herd|放牧:agent={~}}}

4 Conclusion and Future Work

To solve the problem between string processing and conceptual processing, we propose a universal concept representation mechanism, called E-HowNet. Due to the advantage of easy semantic composition and decomposition properties of the model, we are able to compose necessary information for unknown words, phrases and even sentences in text processing. We intend to achieve that any concept can be defined by E-HowNet.

However, there are still some problems that have to further investigate. Discovering semantic structures needs more efforts in both analogical processing and fine-grain features representation. Our hope is to perform automatic detection and assignment of semantic features for sense representations. It will contribute to ontology building and improve the performance of NLP applications.

Acknowledgments. This research was supported in part by National Science Council under a Center Excellence Grant NSC 94-2752-E-001-001-PAE and Grant NSC94-2213-E-001-019.

References

[1] Chen, K. J., Huang, C. R., Chang, L. P., Hsu, H. L.: SINICA CORPUS: Design Methodology for Balanced Corpora. Proceedings of PACLIC 11th Conference (1996) 167-176
[2] Dong, Z. D., Dong, Q.: HowNet. http://www.keenage.com/
[3] Fellbaum, C.: WordNet-An Electronic Lexical Database, the MIT Press. (1998)
[4] Pustejovsky, J.: The Generative Lexicon, Cambridge, MA, The MIT Press (1995).
[5] Fillmore, C.: FrameNet. http://www.icsi.berkeley.edu/~framenet/
[6] Chen, K. J., Huang, S. L., Shih, Y. Y., Chen, Y. J.: Multi-level Definitions and Complex Relations in Extended-HowNet. Workshop on Chinese Lexical Semantics, Beijing University (in Chinese)(2004).
[7] Chen K. J., Huang S. L., Shih Y. Y., Chen Y. J.: Extended-HowNet- A Representational Framework for Concepts, OntoLex 2005 - Ontologies and Lexical Resources IJCNLP-05 Workshop, Jeju Island, South Korea (2005)
[8] Veale, T.: Analogy Generation with HowNet. In Proceedings of the International Joint Conference on Artificial Intelligence (2005) 1148-1153.
[9] Chen, Keh-Jiann, Chen C. J.,: Automatic Semantic Classification for Chinese Unknown Compound Nouns, Proceedings of Coling 2000 (2000), 173-179.