Chapter 15
Learning Hierarchical Lexical Hyponymy

Jiayu Zhou
Arizona State University, USA

Shi Wang
Chinese Academy of Sciences, China

Cungen Cao
Chinese Academy of Sciences, China

ABSTRACT

Chinese information processing is a critical step toward cognitive linguistic applications like machine translation. Lexical hyponymy relation, which exists in some Eastern languages like Chinese, is a kind of hyponymy that can be directly inferred from the lexical compositions of concepts, and of great importance in ontology learning. However, a key problem is that the lexical hyponymy is so commonsense that it cannot be discovered by any existing acquisition methods. In this paper, we systematically define lexical hyponymy relationship, its linguistic features and propose a computational approach to semi-automatically learn hierarchical lexical hyponymy relations from a large-scale concept set, instead of analyzing lexical structures of concepts. Our novel approach discovered lexical hyponymy relation by examining statistic features in a Common Suffix Tree. The experimental results show that our approach can correctly discover most lexical hyponymy relations in a given large-scale concept set.

1. INTRODUCTION

With the advancement of modern information technology, we are facing the increasing need of language processing technologies in Eastern languages like Chinese. During past centuries, ontology building becomes an important part in semantic-level linguistic and knowledge processing, and meanwhile, since most of our knowledge is incarnated within free text as the form of natural language, development of linguistic processing is helping to better ontology learning. Hyponymy relations play an important role in knowledge engineering and the acquisition of which becomes an essential and crucial problem. The hierarchy structure of hyponymy relations composes the

DOI: 10.4018/978-1-4666-1743-8.ch015

skeleton of knowledge bases and application of which ranges natural language processing, information retrieval, machine translation to other related domains.

Several knowledge sources are used in hyponymy acquisition, three primary types of which are: structured corpus (De Meo, Terracina, Quattrone & Ursino, 2004), semi-structured corpus (Dolan, Vanderwende & Richardson, 1993) and unstructured corpus (Cao & Shi, 2001). The largest among the three is unstructured text, the research of which has attracted a lot of researchers and has become a key research area. Thanks to recent research effort on knowledge engineering, new knowledge sources, such as large scale Chinese concept set extracted from unstructured corpus (Wang, Cao, Cao and Cao, 2007; Zhou, Wang & Cao, 2007), are available and have provided rich information.

There are three mainstream approaches—the *Symbolic* approach, the *Statistical* approach and the *Hierarchical* approach—to discovery general hyponymy relations automatically or semi automatically (Du & Li, 2006). The Symbolic approach, depending on lexicon-syntactic patterns, is currently the most popular technique (Hearst, 1992; Liu, Cao, Wang & Chen, 2006; Liu, Cao & Wang, 2005; Ando, Sekine & Ishizaki, 2003). Hearst (1992) was one of the early researchers to extract hyponymy relations from Grolier's Encyclopedia by matching 4 given lexicon-syntactic patterns, and more importantly, she discussed about extracting lexicon-syntactic patterns by existing hyponymy relations. Liu (2005, 2006) used the "isa" pattern to extract Chinese hyponymy relations from unstructured Web corpus, and have been proven to have a promising performance. Zhang (2007) proposed a method to automatically extract hyponymy from Chinese domain-specific free text by three symbolic learning methods. The statistical approach usually adopts clustering and associative rules. Zelenko, Aone and Richardella (2003) introduced an application of kernel methods to extract two certain kinds of hyponymy relations with promising results, combining Support Vector Machine and Voted Perception learning algorithms. The hierarchical approach is trying to build a hierarchical structure of hyponymy relations. Caraballo (1999) built a hypernymy hierarchy of nouns via a bottom-up hierarchical clustering technique, which was akin to manually constructed hierarchy in WordNet.

In this study, we have found a special kind of Chinese hyponymy relationship, called *lexical hyponymy*, which is of great importance in ontology learning. To the best of our knowledge, no existing method can extract these hyponym relations. We propose a semi-automatic lexical hyponymy acquisition approach within a large-scale concept set, which integrates symbolic, statistical and hierarchical techniques. The fundamental process of our method is that we firstly hierarchically cluster the large-scale concept set according to our *common suffix tree clustering* algorithm, we then use some crucial statistics of concept set to construct *suffix concept identification rules*, which are used to extract *class concept candidates*. Then we apply a Google-based *class concept verification* on the candidates. After that, we use a prefix clustering in order to improve the efficiency of the human-involved judgment. Finally, we can export hierarchical lexical hyponymy relations from the common suffix tree. Experimental results prove that our approach can find lexical hyponymy with a promising result.

Recent advances in cognitive informatics and computing witness the growing demand of concepts, and abstract computation on concepts becomes the focuses of current researches in various fields. Concept algebra has been a fundamental part of the Denotational Mathematics framework (Wang, 2008), where concepts serve as a type of new mathematical entities. Success applications of concept algebra include internal knowledge representation, autonomous machine learning, intelligent search engines, etc. (Wang, 2009). Lexical hyponymy described in this paper can be used to extend the relationships of current knowledge