The Semantic Knowledge-base of Contemporary Chinese and its Applications in WSD∗

Hui Wang
Institute of Computational Linguistics
Peking University
Beijing 100871, China
whui@pku.edu.cn

Shiwen Yu
Institute of Computational Linguistics
Peking University
Beijing 100871, China
yusw@pku.edu.cn

Abstract

The Semantic Knowledge-base of Contemporary Chinese (SKCC) is a large scale Chinese semantic resource developed by the Institute of Computational Linguistics of Peking University. It provides a large amount of semantic information such as semantic hierarchy and collocation features for 66,539 Chinese words and their English counterparts. Its POS and semantic classification represent the latest progress in Chinese linguistics and language engineering. The descriptions of semantic attributes are fairly thorough, comprehensive and authoritative. The paper introduces the outline of SKCC, and indicates that it is effective for word sense disambiguation in MT applications and is likely to be important for general Chinese language processing.

Key words: Semantic knowledge-base, lexical semantic, computational lexicography, word sense disambiguation (WSD), Chinese language processing

1 Introduction

Semantic resources play an important role in many areas of Natural Language Processing (NLP). The Institute of Computational Linguistics (ICL) of Peking University has been engaged in research and development of the Semantic Knowledge-base of Contemporary Chinese (SKCC) in the last eight years. This lexicon-building project was a collaboration with the Institute of Computing Technology, Chinese Academy of Sciences during 1994-1998, and resulted in a machine-readable bilingual lexicon suitable for use with Machine Translation applications, which contained a fairly complete characterization of the semantic classification, valence specifications and collocation properties for 49 thousands Chinese words and their English counterparts (Wang Hui, 1998).

Since 2001, the further development of SKCC has been co-conducted by ICL and Chinese Department of Peking University. At present, SKCC has made great progress. Not only is the scale extended to 66,539 entries, but also the quality has been immensely improved. The semantic classification in the updated edition of SKCC is the embodiment of the very latest progress in Chinese linguistics and language engineering, while the semantic descriptions are comprehensive and thorough. It can provide rich lexical semantic information for various NLP applications.

2 Outline of SKCC

2.1 Scale and Structure

SKCC consists of one general database and six sub-databases. The general database contains shared attributes of all the 66,539 entries, while the sub-databases provide detailed descriptions of the

∗ Supported by China National Fundamental Research Program (973) (PN: G1998030507-4).
distinctive semantic attributes associated with the parts of speech (POS). For example, the verb database has 16 attribute fields, noun database and adjective database has 15 attribute fields respectively.

<table>
<thead>
<tr>
<th>Database Name</th>
<th>Entries</th>
<th>Attribute fields</th>
<th>Attribute value</th>
</tr>
</thead>
<tbody>
<tr>
<td>nouns</td>
<td>38,478</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>verbs</td>
<td>21,142</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>adjective</td>
<td>5,577</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>pronouns</td>
<td>236</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>adverbs</td>
<td>997</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>numerals</td>
<td>109</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>General</td>
<td>66,539</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>133,071</td>
<td><strong>91</strong></td>
<td><strong>1</strong></td>
</tr>
</tbody>
</table>

Table 1: Scale of SKCC

All of the six sub-databases can be linked to the general database through four key fields, namely ENTRY, POS, HOMOMORPHISM and SENSE. As a result, the son knots can inherit all information from their father knots (Figure 1).

![Figure 1 Main structure of SKCC](image)

2.2 Semantic Hierarchy

One of the most outstanding characteristics of SKCC is that its semantic hierarchy is based on grammatical analysis, rather than merely on general knowledge (as illustrated in Figure 2 below). This classification system represents the latest progress in Chinese semantics. It is very useful for NLP applications (Zhan Weidong, 1997), as well as compatible with various semantic resources, such as Wordnet (Christiane Fellbaum, 1998), Chinese concept dictionary (CCD) (Yu Jiangsheng, 2002), HowNet (Dong Zhendong, 2000) etc. Currently, the classification of all of the 66,539 entries has already been completed.