Introduction

It is a great pleasure for me to present the second issue from the third Chinese Lexical Semantics Workshop (CLSW3). CLSW3 was held from May 1-4, 2002 at Academia Sinica in Taipei. The first two workshops (also known as Chinese Language Science Workshop: Lexical Semantics) were held at the Institute of Chinese Linguistics, City University of Hong Kong in October 2000 and at the Institute of Computational Linguistics, Peking University in May 2001 respectively. The next workshop (CLSW4) returns to the City University of Hong Kong in 2003.

CLSW3 shares with its predecessors an interdisciplinary approach as well as a focus on how meaning is represented and processed in Chinese. The versatility of the research approaches employed made it necessary for the papers to be published in two different fields to ensure high-quality peer review.

Eight of the papers from this workshop were published in *Computational Linguistics and Chinese Language Processing*. Six of them appeared in Vol.7 No.2 (papers 3, 4, 7, 11, 13, and 15. A table follows this introduction with a list of all the papers in alphabetical order. The numbers listed here refer to that table.). Two more will appear in Vol.8 No.2 (2 and 8). The remaining 8 papers (1, 5, 6, 9, 10, 12, 14, and 17) appear in the current volume (*Language and Linguistics*, Vol.4, No.3). In addition, we were able to include an additional paper that was submitted independently to *Language and Linguistics* in this volume because of its relevance to this special issue. This is paper number 16 by Yu and Hu.

It is crucial, in spite of the physical separateness in two different journals, to recognize that these papers are unified by several fundamental issues involving meaning in the Chinese lexicon. Note that the interdisciplinary nature of the work reported means that many papers deal with more than one issue. These issues focus on the following:

1. **How is meaning lexically represented in Chinese?**
2. **Are there cross-lingual structural correspondences in lexical semantics and how can cross-lingual lexical semantics help Chinese processing?**
3. **What are applications for Chinese semantic processing?**

The first issue involves both linguistic and computational lexical semantics, as well as (computational) lexicography. Among the papers, Ahrens et al. (1), Gao and Cheng (5), Lai (9), Liu (10), Shen (12), and Zhan (17) share a linguistic lexical semantic base, although several different theories are adopted. Following the tradition of computational lexical semantics are Chen and You (3), Chen et al. (4), Ker (8), Liu and Li (11). Lastly, Kang (7), Yu and Hu (16) provide a perspective from computational lexicography.

The foci on the second issue include accounting for bilingual lexical semantic comparisons: Gao and Cheng (5), Ahrens et al. (1), and T'sou and Kwong (14). One paper accounts of structural correspondences: Huang et al. (6). Lastly, three papers deal with bilingual information with respect to Chinese lexical semantic processing. They are Chang et al. (2), Huang et al. (6), and Ker (8).

The last issue is computation-oriented. The computational goals include semantic similarity, as in Chen and You (3) and Liu and Li (11), semantic networks (such as WordNet), as in Chang et al. (2), Huang et al. (6), and Ker (8), dealing with new words or new information, as in Song and Xu (13), and Yu and Hu (16), and lastly dealing with disambiguation, as in Wang (15).

In order to keep the integrity of the special issues and to provide easier cross-references, we include in this volume the abstracts for all the papers that appeared in *CLCLP*.

The two special issues from CLSW3 are indebted to the guidance and support of the chief editors of the two journals: Dah-an Ho for *Language and Linguistics*, and Keh-Jiann Chen for *CLCLP*. We would like to thank all authors for their contributions, both at the workshop and to these special volumes. We would also like to express our gratitude for the timely reviews by all the reviewers. Last, but not least, we thank the editorial assistants at both journals for working tirelessly on these volumes.

Chu-Ren Huang
May 25, 2003, on behalf of

Editorial Committee of the CLSW3 Special Issues
Chen, Keh-Jiann. Institute of Information Science, Academia Sinica
Cheng, Chin-Chuan. Institute of Chinese Linguistics, City University of Hong Kong
Huang, Chu-Ren. Institute of Linguistics, Academia Sinica
Su, Lily I-Wen. Graduate Institute of Linguistics, National Taiwan University
Sun, Maosong. Department of Computer Science, Tsing Hua University
T'sou, Benjamin K. Language Information Science Research Centre, City University of Hong Kong
Tseng, Shu-Chuan. Institute of Linguistics, Academia Sinica
Yu, Shiwen. Institute of Computational Linguistics, Peking University

Table of the papers that appear in the two volumes listed in alphabetical order by the author’s last name.

No	Author	Title	Journal Issue
1	Kathleen Ahrens, Chu-Ren Huang and Yuan-hsun Chuang	Sense and Meaning Facets in Verbal Semantics: A MARVS Perspective	L&L Vol.4 No.3
2	Jason S. Chang et al.	Building A Chinese WordNet Via Class-Based Translation Model	CLCLP Vol.8 No.2
3	Keh-Jiann Chen and Jia-Ming You	A Study on Word Similarity Using Context Vector Models	CLCLP Vol.7 No.2
4	陳祖舜、周強、趙強	情境——組織/存放辭彙語義知識的恰當框架	CLCLP Vol.7 No.2
5	Hong Gao and Chin-Chuan Cheng	Verbs of Contact by Impact in English and Their Equivalents in Mandarin Chinese	L&L Vol.4 No.3
6	Chu-Ren Huang, Elanna I. J. Tseng, Dylan B. S. Tsai, Brian Murphy	Cross-lingual Portability of Semantic Relations: Bootstrapping Chinese WordNet with English WordNet Relations	L&L Vol.4 No.3
7	亢世勇	《現代漢語新詞語資訊電子詞典》的研究與實現	CLCLP Vol.7 No.2
8	柯淑津	從詞網出發的中文複合名詞的語意表達	CLCLP Vol.8 No.2
9	Huei-ling Lai	The Semantic Extension of Hakka LAU	L&L Vol.4 No.3
10	Mei-chun Liu	From Collocation to Event Information: The Case of Mandarin Verbs of Discussion	L&L Vol.4 No.3
11	劉群、李素建	基於《知網》的辭彙語義相似度計算	CLCLP Vol.7 No.2
12	沈陽	“V 著 A”結構分化的詞匯語義條件	L&L Vol.4 No.3
13	宋柔、許勇	基於詞彙語義的百科辭典知識提取實驗	CLCLP Vol.7 No.2
14	Benjamin K. T’sou and Oi Yee Kwong	When Laws Get Common: Comparing the Use of Legal Terms in Two Corpora	L&L Vol.4 No.3
15	王惠	基於組合特徵的漢語名詞詞義消歧	CLCLP Vol.7 No.2
16	俞士汶、胡俊峰	唐宋詩之詞匯自動分析及應用	L&L Vol.4 No.3
17	詹衛東	漢語述結式的組配約束及“v＋a＋n”歧義格式分析	L&L Vol.4 No.3

Building A Chinese WordNet Via Class-Based Translation Model

Jason S. Chang　Tracy Lin　Geeng-Neng You
Thomas C. Chuang　Ching-Ting Hsieh

Semantic lexicons are indispensable to research in lexical semantics and word sense disambiguation (WSD). For the study of WSD for English text, researchers have been using different kinds of lexicographic resources, including machine readable dictionaries (MRDs), machine readable thesauri, and bilingual corpora. In recent years, WordNet has become the most widely used resource for the study of WSD and lexical semantics in general. This paper describes the Class-Based Translation Model and its application in assigning translations to nominal senses in WordNet in order to build a prototype Chinese WordNet. Experiments and evaluations show that the proposed approach can potentially be adopted to speed up the construction of WordNet for Chinese and other languages.

(*CLCLP* Vol.8 No.2)

A Study on Word Similarity Using Context Vector Models

Keh-Jiann Chen　　Jia-Ming You

There is a need to measure word similarity when processing natural languages, especially when using generalization, classification, or example-based approaches. Usually, measures of similarity between two words are defined according to the distance between their semantic classes in a semantic taxonomy. The taxonomy approaches are more or less semantic-based that do not consider syntactic similarities. However, in real applications, both semantic and syntactic similarities are required and weighted differently. Word similarity based on context vectors is a mixture of syntactic and semantic similarities.

In this paper, we propose using only syntactic related co-occurrences as context vectors and adopt information theoretic models to solve the problems of data sparseness and characteristic precision. The probabilistic distribution of co-occurrence context features is derived by parsing the contextual environment of each word, and all the context features are adjusted according to their IDF (inverse document frequency) values. The agglomerative clustering algorithm is applied to group similar words according to their similarity values. It turns out that words with similar syntactic categories and semantic classes are grouped together.

(*CLCLP* Vol.7 No.2, pp.37-58)

情境——組織/存放辭彙語義知識的恰當框架
Situation—A Suitable Framework for Organizing and Positioning Lexical Semantic Knowledge

陳祖舜 (Zusun Chen) 周 強 (Qiang Zhou) 趙 強 (Qiang Zhao)

The characteristic and an advantage of natural language is that, as a symbolic system, it has an internal logical framework for organizing and positioning conceptual knowledge, which is its lexicon system. This framework implements the fundamental function of natural language to condense, absorb, organize and position conceptual knowledge, and creates progressively a very large and complex build-in knowledge system in the language. It is also the basis of the other two fundamental functions of natural language; i.e., it serves as a tool for communication and as a medium for conceptual thought. The natural language semantics should reproduce the basic framework of natural language in their theoretic realms to represent these three functions and their relationships. The lexical semantics thereby become their core.

A word is the symbolic embodiment of a concept, and a concept is generated in a peculiar cognition scheme, which will be called its generating scheme. We cannot describe and define a concept clearly unless we put it into its generating scheme. Meanwhile, the implementation of the concept involves a procedure that contrasts, restores, and refers to its generating scheme in a special environment, which will be called its application scheme.

We propose to use the situation as a mathematical model to describe a cognition scheme. Therefore, the situation theory can serve as a unified theoretical framework for constructing the lexical semantics and the natural language semantics built upon it, as mentioned above. Therefore, many new viewpoints are proposed. In this paper, only some elementary questions about them are discussed, including: 1) using a situation to express a scheme and using a situation to describe a concept (this is the key point of the paper); 2) formulating the situation algebra for describing relations, transformations, and operations for situations so as to simulate conceptual thinking by means of algebraic calculus; 3) constructing a situation network to implement a scheme structure and conceptual structure, where the key point is the constitution and organization of a semantic dictionary. We use some practical cases to illustrate these methods. The mathematical theory relevant to them will be presented in our future papers.

Key words: concept, lexical meaning, situation, situation algebra, semantic dictionary, lexical semantics

(*CLCLP* Vol.7 No.2, pp.1-36)

《現代漢語新詞語資訊電子詞典》的研究與實現
Development and Study of the "Modern Chinese New Words Information Electronic Dictionary"

亢世勇 (Shiyong Kang)

We introduce the development of the Electronic Lexicon of Contemporary Newborn Chinese Words: (1) the definition of a newborn word, (2) the main principle behind constructing the lexicon, (3) the collection of newborn words and their feature descriptions of them, and (4) the classification of 40,000 newborn words. In our opinion, a new bornword is a character string that appeared after 1978 in a new form, with a new meaning and with a new usage. In addition, it must be frequently used and accepted, but the names of men and places are not newborn words according to our definition. The approach to collecting newborn words is quite unrestricted, that is, the more the better. Based on the Contemporary Chinese Grammatical Knowledge Base of the Institute of Computational Linguistics at Peking University, we have finished compiling a lexicon of almost 40,000 newborn words semi-automatically. The lexicon, we believe, is a worthy resource for research on Chinese word-building rules and Natural Language Processing. Firstly, classification is done based on the preponderant grammatical characteristics of each word, and then the detailed features are described in the database of ACCESS. The lexicon contains a total base and three grammatical bases (i.e., a noun base, verb base and adjective base); what's more, it also has an old word base, a loanword base and a acronym base. The entire base is related to the sub-bases through the fields of word, phonetic notation and semantics fields, which form a hypernymy hierarchy that is quite convenient for searching. Totally, there are more than 200 fields in the bases that give information regarding phonetic notation, semantics, sources, word building, syntax and pragmatics. Without doubt, this lexicon is one of the largest domestic lexicons available with the most detailed descriptions of newborn Chinese words.

Key words: Chinese information processing, new words, electronic dictionary

(*CLCLP* Vol.7 No.2, pp.89-100)

從詞網出發的中文複合名詞的語意表達

柯淑津 (Sue-Jin Ker)

WordNet provides plenty of lexical meaning; therefore, it is very helpful in natural language processing research. Each lexical meaning in Princeton WordNet is presented in English. In this work, we attempt to use a bilingual dictionary as the backbone to automatically map English WordNet to a Chinese form. However, we encounter many barriers between the two different languages when we observe the preliminary result for the linkage between English WordNet and the bilingual dictionary. This mapping causes the Chinese translation of the English synonym collection (Synset) to correspond to unstructured Chinese compound words, phrases, and even long string sentence instead of independent Chinese lexical words. This phenomenon violates the aim of Chinese WordNet to take the lexical word as the basic component. Therefore, this research will perform further processing to study this phenomenon.

The objectives of this paper are as follows: First, we will discover core lexical words and characteristic words from Chinese compound words. Next, those lexical words will be expressed by means of conceptual representations. For the core lexical words, we use grammar structure analysis to locate such words. For characteristic words, we use sememes in HowNet to represent their lexical meanings. Certainly, there exists a problem of ambiguity when Chinese lexical words are translated into their lexical meanings. To resolve this problem, we use lexical parts-of-speech and hypernyms of WordNet to reduce the lexical ambiguity. We experimented on nouns, and the experimental results show that sense disambiguation could achieve a 93.8% applicability rate and a 93.5% correct rate.

(*CLCLP* Vol.8 No.2)

基於《知網》的辭彙語義相似度計算
Word Similarity Computing Based on How-net

劉　群 (Qun Liu)　　李素建 (Sujian Li)

Word similarity is broadly used in many applications, such as information retrieval, information extraction, text classification, word sense disambiguation, example-based machine translation, etc. There are two different methods used to compute similarity: one is based on ontology or a semantic taxonomy; the other is based on collocations of words in a corpus.

As a lexical knowledgebase with rich semantic information, How-net has been employed in various researches. Unlike other thesauri, such as WordNet and Tongyici Cilin, in which word similarity is defined based on the distance between words in a semantic taxonomy tree, How-net defines a word in a complicated multi-dimensional knowledge description language. As a result, a series of problems arise in the process of word similarity computation using How-net. The difficulties are outlined below:

1. The description of each word consists of a group of sememes. For example, the Chinese word “暗箱 (camera obscura)” is described as: “part|部件, #TakePicture|拍攝, %tool|用具, body|身”, and the Chinese word “寫信 (write a letter)” is described as: “write|寫, ContentProduct =letter|信件”;

2. The meaning of a word is not a simple combination of these sememes. Sememes are organized using a specific knowledge description language.

To meet these challenges, our work includes:

1. A study on the How-net knowledge description language. We rewrite the How-net definition of a word in a more structural format, using the abstract data structure of *set* and *feature structure*.

2. A study on the algorithm used to compute word similarity based on How-net. The similarity between sememes, that between *sets*, and that between *feature structures* are given. To compute the similarity between two sememes, we use the distance between the sememes in the semantic taxonomy, as is done in Wordnet and Tongyici Cilin. To compute the similarity between two *sets* or two *feature structures*, we first establish a one-to-one mapping between the elements of the *sets* or the *feature structures*. Then, the similarity between the *sets* or *feature structures* is defined as the weighted average of the similarity between their elements. For *feature structures*, a one-to-one mapping is established according to the attributes. For *sets*, a one-to-one mapping is established according to the similarity between their elements.

3. Finally, we give experiment results to show the validity of the algorithm and compare them with results obtained using other algorithms. Our results for word similarity agree with people’s intuition to a large extent, and they are better than the results of two comparative experiments.

Key words: How-net, word similarity computing, natural language processing

(*CLCLP* Vol.7 No.2, pp.59-76)

基於詞彙語義的百科辭典知識提取實驗

An Experiment on Knowledge Extraction from an Encyclopedia Based on Lexicon Semantics

宋　柔 (Rou Song)　　許　勇 (Yong Xu)

The typical approaches to extracting text knowledge are sentential parsing and pattern matching. Theoretically, text knowledge extraction should be based on complete understanding, so the technology of sentential parsing is used in the field. However, the fragility of systems and highly ambiguous parse results are serious problems. On the other hand, by avoiding thorough parsing, pattern matching becomes highly efficient. However, different expressions of the same information will dramatically increase the number of patterns and nullify the simplicity of the approach.

Parsing in Chinese encounters greater barriers than that in English does. Firstly, Chinese lacks morphology. For example, recognition of base-NP in Chinese is more difficult than that in English because its left boundary is hard to discern. Secondly, there are many stream sentences in Chinese which lack subjects and cause parsing to fail. Finally, in Chinese, the absence of verbs is also pervasive. Sentential parsing centering on verbs, which is used with English, is not always successful with Chinese.

We are engaged in research on knowledge extraction from the Electronic Chinese Great Encyclopedia. Our goal is to extract unstructured knowledge from it and to generate a well-structured database so as to provide information services to users. The pattern-matching approach is adopted.

The experiment was divided into two steps: (1) classifying entries based on lexicon semantics; (2) establishing a formal system based on lexicon semantics and extracting knowledge by means of pattern matching.

Classification of entries is important because in the text of the entries of different categories there are different kinds of patterns expressing knowledge. Our experiment demonstrated that an entry of the encyclopedia can be classified precisely merely according to the characters in the entry and the words in the first sentence of the entry's text. Some specific categories, e.g., organization names and Chinese place names, can be classified satisfactorily merely according to the suffix of the entry, for suffixes are closely related with semantic categories in Chinese.

The formal system designed for knowledge extraction consists of 4 kinds of meta knowledge: concepts, mapping, relations and rules, which reflect lexicon semantic attributes. The present experiment focused on the extraction of knowledge about various areas from the texts regarding administrative places of China (how large is a place or its subdivisions). The results of the experiment show that the design of the formal system is practical. It can accurately and completely denote various expressions of simple knowledge in a Chinese encyclopedia. However, when the focus of knowledge changes, e.g., from administrative areas to habits of animals, it is a labor-intensive task to renew the formal system. Therefore the study of auto or semi-auto generation of this kind of formal system is required.

(*CLCLP* Vol.7 No.2, pp.101-112)

基於組合特徵的漢語名詞詞義消歧
A Study on Noun Sense Disambiguation Based on Syntagmatic Features

王　惠 (Hui Wang)

Word sense disambiguation (WSD) plays an important role in many areas of natural language processing, such as machine translation, information retrieval, sentence analysis, and speech recognition. Research on WSD has great theoretical and practical significance. The main purposes of this study were to study the kind of knowledge that is useful for WSD, and to establish a new WSD model based on syntagmatic features, which can be used to disambiguate noun sense in Mandarin Chinese effectively.

Close correlation has been found between lexical meaning and its distribution. According to a study in the field of cognitive science (Choueka, 1983), people often disambiguate word sense using only a few other words in a given context (frequently only one additional word). Thus, the relationships between one word and others can be effectively used to resolve ambiguity. Based on a descriptive study of more than 4,000 Chinese noun senses, a multi-level framework of syntagmatic analysis was designed to describe the syntactic and semantic constraints of Chinese nouns. All of these polyseme nouns were surveyed, and it was found that different senses have different and complementary distributions at the syntax and/or collocation levels. This served as a foundation for establishing an WSD model by using grammatical information and a thesaurus provided by linguists.

The model uses *the Grammatical Knowledge-base of Contemporary Chinese* (Yu Shiwen et al. 2002) as one of its main machine-readable dictionaries (MRDs). It can provide rich grammatical information for disambiguation of Chinese lexicons, such as parts-of-speech (POS) and syntax functions.

Another resource of the model is *the Semantic Dictionary of Contemporary Chinese* (Wang Hui et al. 1998), which provides a thesaurus and semantic collocation information of more than 20,000 nouns. They were employed to analyze 635 Chinese polysemous nouns.

By making full use of these two MRD resources and a very large POS-tagged corpus of Mandarin Chinese, a multi-level WSD model based on syntagmatic features was developed. The experiment described at the end of the paper verifies that the approach achieves high levels of efficiency and precision.

Key words: word sense disambiguation, syntagmatic features, noun sense, Chinese language information processing

(*CLCLP* Vol.7 No.2, pp.77-88)