Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 417-430

The Sinica Sense Management System: Design and Implementation

Chu-Ren Huang*, Chun-Ling Chen*, Cui-Xia Weng*, Hsiang-Ping Lee*,

Yong-Xiang Chen* and Keh-Jiann Chen*

Abstract

A sense-based lexical knowledgebase is a core foundation for language engineering. Two important criteria must be satisfied when constructing a knowledgebase: linguistic felicity and data cohesion. In this paper, we discuss how data cohesion of the sense information collected using the Sinica Sense Management System (SSMS) can be achieved. SSMS manages both lexical entries and word senses, and has been designed and implemented by the Chinese Wordnet Team at Academia Sinica. SSMS contains all the basic information that can be merged with the future Chinese Wordnet. In addition to senses and meaning facets, SSMS also includes the following information: POS, example sentences, corresponding English synset(s) from Princeton WordNet, and lexical semantic relations, such as synonym/antonym and hypernym/hyponym. Moreover, the overarching structure of the system is managed by using a sense serial number, and an inter-entry structure is established by means of cross-references among synsets and homographs. SSMS is not only a versatile development tool and management system for a sense-based lexical knowledgebase. It can also serve as the database backend for both Chinese Wordnet and any sense-based applications for Chinese language processing.

Keywords: Lexical Knowledgebase, Word Senses, Chinese Wordnet, Lexical Semantic Relation

* Institute of Linguistics, Academia Sinica, 128, Section 2, Academia Road, Nankang, Taipei, 115, Taiwan Tel: 886-2-26523108 Fax: 886-2-27856622
E-mail: churen@gate.sinica.edu.tw

[Received April 10, 2005; Revised June 16, 2005; Accepted August 20, 2005]

1. Background and Motivation

A sense-based lexical knowledgebase is a core foundation for language engineering. WordNet and Euro WordNet are two well-known examples. Two important criteria must be satisfied when constructing a knowledgebase: linguistic felicity and data cohesion. Huang *et al.* discussed how linguistic felicity can be achieved when building a comprehensive inventory of Chinese senses from corpus data [Huang *et al.* 2003]. They introduced five criteria as well as operational guidelines for sense distinction. In this paper, we will discuss how data cohesion of the sense information collected using the Sinica Sense Management System (SSMS) can be achieved.

2. Introduction to the Content of the SSMS

The structure of a lexical semantic web can provide not only materials for linguistic research but also an infrastructure for NLP and other applications. Two main tasks are involved in constructing a Wordnet. One is distinguishing synsets which are clustered according to word senses, and the other one is collecting the semantic relations connecting the synsets. A wordnet is formed by taking these synsets as the nodes and then connecting each node by using the semantic relations. Of the two tasks, constructing the synsets is the most fundamental work. Constructing a synset involves classifying a series of synonyms that carry one specified lexical concept. A polysemous word is assigned more than one synset in order to show the variety of word senses. Analyzing and distinguishing word senses are two crucial steps when constructing a Wordnet. In order to lay a solid theoretical foundation of the current work, we developed a series of principles for analyzing Chinese word senses [Huang *et al.* 2003]. These principles not only satisfy the conditions of felicity and cohesion but also serve as guidelines for distinguishing and analyzing large numbers of Chinese word senses. Based on the word-sense identification criteria, the linguistic knowledge can be consistently described, and we can easily map the linguistic knowledge to ontologies or transform it into formal representations.

The SSMS system is designed to store and manage word sense data generated in the analysis stage. SSMS manages both lexical entries and word senses. This system has been designed and implemented by the Chinese Wordnet Team at Academia Sinica. It contains all the basic information that can be merged with the future Chinese Wordnet. SSMS is meaning-driven. Each sense of a lemma is identified specifically and given a separate entry. When further differentiation at the meaning facet level is called for, each facet of a sense is also described in a full entry [Ahrens *et al.* 1998]. In addition to senses and meaning facets, this system also includes the following information: POS, example sentences, corresponding English synset(s) from Princeton WordNet, and lexical semantic relations, such as synonym/antonym and hypernym/hyponym. Moreover, the overarching structure of the

system is managed using a sense serial number, and the inter-entry structure is established by means of cross-references among synsets and homographs.

Currently, the Chinese Wordnet Team is focusing on analyzing middle-frequency words in the Sinica Corpus. Our reason for choosing middle-frequency words as our target ones is that with only three to five senses of a word, we can investigate the senses and meaning facets of each word deeply and accurately, while avoiding the simple situation of one sense for low-frequency words and the complicated situation of numerous senses for high-frequent words. So far, nearly 3,000 more lemmas have been analyzed, and close to 4,000 senses have been identified. These data are presented in a technical report that is updated yearly [Huang *et al.* 2005]. In the near future, these results will be used as a basis for Natural Language Processing or E-learning applications.

3. Design Principle of SSMS

A sense-based lexical knowledgebase with data cohesion must meet three requirements: unique identification of senses, trackability of senses, and consistent sense definitions. SSMS uses four tools to satisfy these requirements.

3.1 Unique Serial Number

Each sense or meaning facet is identified by a unique serial number in SSMS. In Princeton WordNet [Fellbaum 1998], each synset is given a unique offset number. However, the offset number does not have a logical structure. Hence, although it guarantees unique identification, it is not easily trackable. An alternative approach is to set up a base ontology and assign senses to an ontological node with a unique ID. However, this is not feasible since we cannot pre-designate all the possible conceptual and semantic relations. In addition, if a decision is made to encode only certain higher level nodes, the random assignment problem will occur because of the coarse granularity inherent to an upper ontology. In our system, the unique serial number of each sense is composed of three segments: sequential information indicating when the lemma was processed, the lemma form, and the sense classification code for each lemma (including the meaning facet level). Take “bao4 zhi3 (newspaper)” as an example: “bao4 zhi3” has two senses as well as two meaning facets for the first sense. The lexical entry of “bao4 zhi3” is as follows:

Example 3-1: The result of sense distinction for “bao4 zhi3 (newspaper)”

報紙 bao4 zhi3 ㄅㄠˋ ㄓˇ

詞義 1：【名詞，Na】指定期出版，報導新聞、提供各式訊息的出版品。

義面 1：指刊物，尤其指內容部份。{ newspaper, 03039218N }

例句：儘管他出現在**報紙**頭條的頻率極高，被刊登的卻幾乎都是片段性的談話。

義面 2：指定期出版，報導新聞、提供各式訊息的紙張本身。{ newspaper, 04738466N }

例句：他找了一張**報紙**，平鋪在面前，取下身邊掛著的匣子之後就開始自言自語。

詞義 2：【名詞，Na】指定期出版，報導新聞、提供各式訊息出版品的組織。{ newspaper, 06009637N }

例句：**報紙**對他進行專訪的內容將刊登於隔天的頭條新聞上。

The four-level unique serial number shown below contains four segments of the unique serial number for the first meaning facet of the first sense of “bao4 zhi3”. Note that the lemma form ID “0018” can be replaced by the actual lemma form “報紙” or “bao4 zhi3” for processing purposes:

	報紙 “bao4 zhi3 (newspaper)”
Lemma processing year	03-
Lemma form ID	-0018-
The first sense	-01-
The first meaning facet	-01

There are four advantages to managing the sense database with unique serial numbers. First, the sequential number not only gives a unique code for each lemma; it also enables a project manager to track work progress more easily. Second, including the lemma in the serial number helps human users to quickly identify the relevant senses. It also facilitates man-machine interaction, such as keyword search for senses. Third, it also provides a logical structure for the sense serial number, since each lemma represents a small number of possible senses. Lastly, four digits are reserved to identify senses and meaning facets belonging to each lemma. The first two digits are reserved for senses and the last two for meaning facets. These four digits also allow the minimal amount of space needed to identify exact sense in the database. For instance, when stipulating a synonym, we can identify it as word0200, which refers to the second sense of a certain lemma. There is no need to repeat the complete sense serial number. The sense serial number enables unique identification and also assists trackability.

3.2 Cross-Reference Device

SSMS automatically prompts all possible cross-references. When a lemma is called up for analysis, all existing records that contain this lemma are prompted. These include either lexical semantic relations, such as synonyms and hyponyms, sense definitions that contains this lemma, or explanatory notes that contain this lemma. In addition to offering rich semantic association information, this feature allows sense relations to be clearly defined, and inconsistencies to be detected. In addition, any anomaly in a definition or expression format can also be discovered. This process also helps us to narrow our focus to a set of control vocabulary for sense definitions. This feature also improves both the trackability of senses and the consistency of sense definitions. For example, if we enter the term “you2 mu4 (nomadic)” in a query, SSMS will automatically prompt two possible relevant references, “man3 (Manchu)” and “meng2 gu3 (Mongol),” both of which refer to areas where nomads live.

3.3 Concurrent Access to the Lexical Knowledgebase and Corpus

In addition, SSMS enables parallel and concurrent access to the lexical knowledgebase and corpus. When a lemma is chosen in the system, all tagged examples of that lemma from the Sinica Corpus are retrieved. This allows closer examination of how the senses are used and distributed. It also enables automatic selection of corpus example sentences. In turn, when sense classification is completed, SSMS allows all the corpus sentences to be sense-tagged and returned to be merged with the original corpus. In other words, a sense-tagged corpus is being processed in parallel. This feature allows each lexical sense to be traced to its actual uses in the corpus. It also allows linguists to examine the data supporting each sense classification. For example, if we enter the term “you2 mu4 (nomadic)” in a query, 9 tagged examples of the lemma from the Sinica Corpus will be retrieved, as shown in *Figure 1.*

Figure 1. Tagged examples of the lemma “you2 mu4 (nomadic)”from Sinica corpus

3.4 Linking to the Sinica BOW

Lastly, SSMS is also linked to the bilingual wordnet information at Sinica BOW. Candidate English synset correspondences, including offset numbers, are shown after a Chinese lemma is chosen. This aids cross-lingual trackability and consistency.

4. The Implementation of SSMS

The implementation of SSMS can be divided into three stages as shown in *Figure 2*. In the lemma analysis stage, based on the criteria and operational guidelines proposed by [Huang 2003], we distinguished senses and meaning facets for each word. At the same time, the Sinica Corpus and Wordnet were referred to for POS, examples, and English translations. Then with the help of dictionary resources or word mapping done by the system, we determined the word relations. The second stage involved two steps. First, we designed the schema of the sense management system database for storing analysis results obtained in the first stage. Then, for the purpose of database management, we developed an interface to help the Chinese Wordnet Team accessing and revising the database. We employed the DELPHI tool to design our system interface. Through the interface, the data in the database can also be exported in lexicon documents. The last stage of this system implementation is the application stage. The aim of our project is to build Chinese Wordnet web sites for user querying. The development languages for these web pages are HTML and ASP. The web pages in these web sites will be viewed through web servers. Through the Internet, people will be able to retrieve data from our sense management database system everywhere, at anytime. The flow of the Sinica Sense Management System is displayed in the following chart. There note that when the initial goals of all three stages are met, work on the first stage will continue to expend the database.

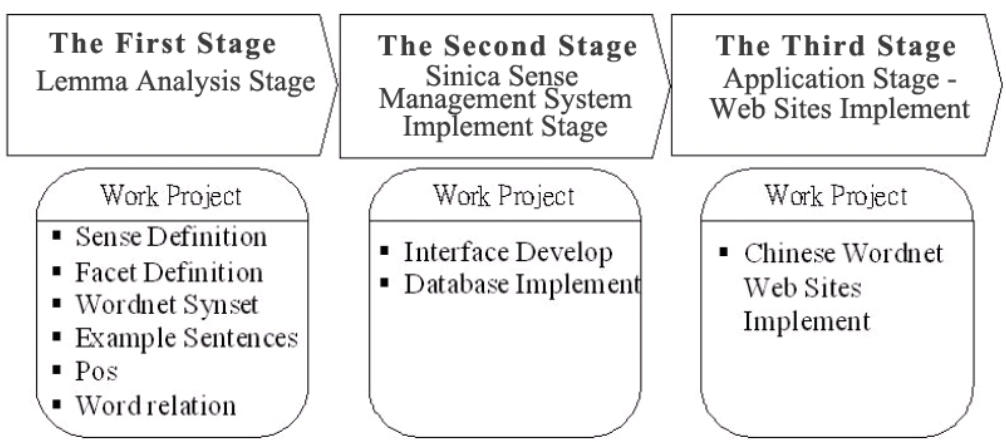

Figure 2. The flow chart of the Sinica Sense Management System

We represent the overall framework of SSMS diagrammatically in *Figure 3.* As the diagram indicates, the Chinese Wordnet Team uses SSMS to access the database and electronic documents as Lexicon reports. Moreover, users on the Internet can browse HTML/ASP pages to query the database through web servers.

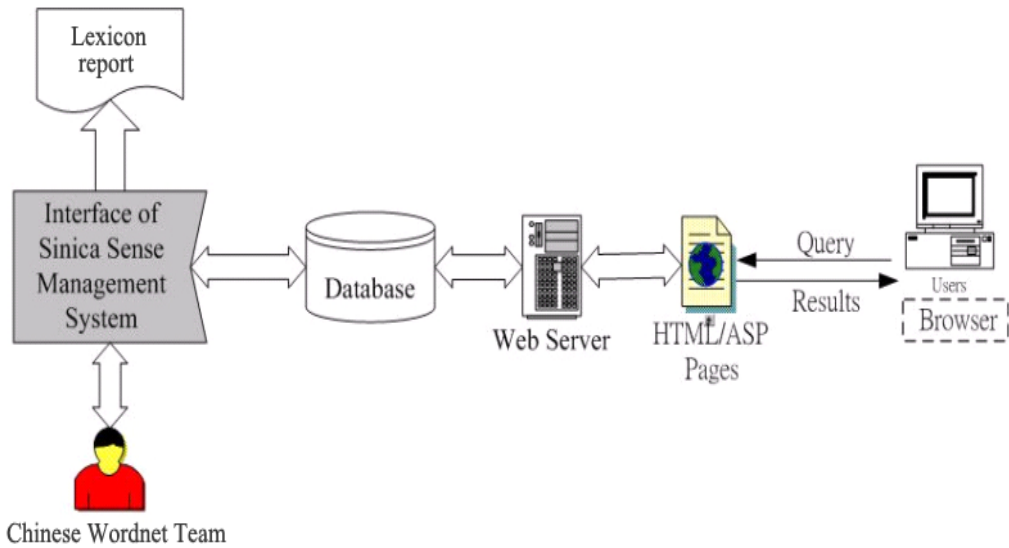

Figure 3. The overall structure of SSMS

4.1 The Schema of the SSMS Database in Class Diagrams

In the section, we will discuss the schema of the SSMS Database. The Unified Modeling Language (UML) [Booch *et al.*1999; Oestereich 1999] is a graphical notation that provides a conceptual foundation for assembling a system out of components from 4+1 views and nine diagrams. Each view is a projection into the organization and structure of the system, focusing on a particular aspect of that system.

We employ the class diagram notations in UML to provide a static view of application concepts in terms of classes and their relationships including generalization and association. Therefore, we will only provide some details about class diagrams in the following.

Class diagrams [Booch *et al.*1999; Oestereich 1999; Mullar 1999] commonly contain the following features:

1. A class diagram shows a set of classes and their relationships. For example, the class diagram of the Suppliers-and-Parts database is shown in *Figure 4.* The terms in italics in *Figure 4* indicate concepts concerning class diagrams.

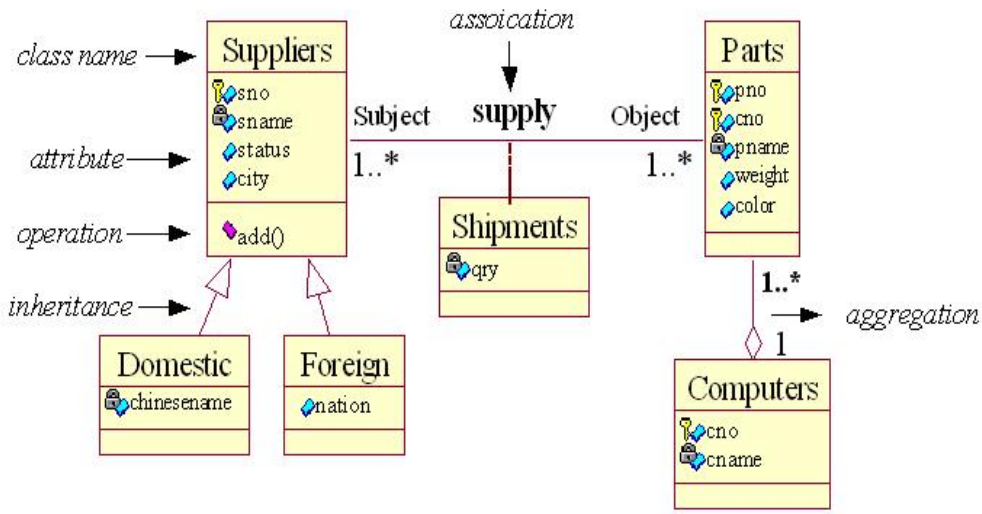

Figure 4. A class diagram for the Suppliers-and- Parts Database

2. A class is a description of a set of objects that share the same attributes, operations, relationships, and semantics. A class mainly contains three important parts: its name, attributes, and operations. We will explain these terms in the following:

(a) Class name: every class must have a name to distinguish it from other classes. For example, **Suppliers** and **Parts** are class names.

(b) Attribute: an attribute represents some property that is shared by all objects of that class. A class may have any number of attributes or no attributes at all. For example, in *Figure 4,* the class **Suppliers** has some attributes, such as *sno, sname, city.*

(c) Operation: an operation is the implementation of a service that can be requested from any object of the class in order to influence behavior. A class may have any number of operations or no operations at all. For example, in *Figure 4,* the class **Suppliers** has one operation: *add().*

3. There are three kinds of relationships between classes:

(a) Association: an association is a structural relationship that specifies objects of one thing to be connected to objects of another. For example, in *Figure 4,* a line drawn between the involved classes (**Suppliers** and **Parts**) represents an association named *supply.*

(b) Aggregation: an aggregation is a "whole/part" relationship, in which one class represents a larger thing (the "whole" class), which consists of smaller things (the "parts" class). Moreover, an aggregation represents a "has-a" relationship, which means that an object of the "whole" class has objects of the "part" class. To represent an aggregation, an empty diamond will be drawn at the "whole" class end of the line linking two classes.

(c) Inheritance: An inheritance relationship can be regarded as a generalization (or specialization), which is a taxonomic relationship between a general (super class) and a special (subclass) element, where the special element adds properties to the general one and behaves in a way that is compatible with it. Therefore, it is sometimes called an "is-a-kind-of" relationship. An inheritance relation is represented by means of a large empty arrow pointing from the subclass to the super class. For example, in *Figure 4*, **Domestic** and **Foreign** suppliers (two subclasses) together are a kind of supplier (the super class).

Based on the need for SSMS content and design principle, *Figure 5* shows the schema of SSMS database using the concepts of class diagrams.

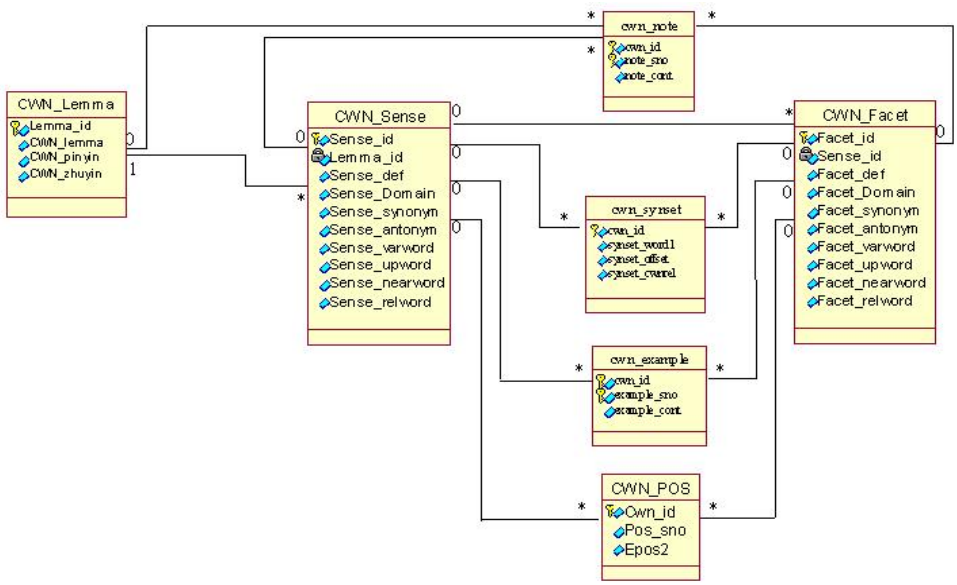

Figure 5. The schema of the Sinica Sense database

4.2 The Function of SSMS

In this section, we will discuss the interface marking for SSMS. Our task orientation is to design a clear and easy interface because we are trying to provide a helpful tool that the members of the Chinese Wordnet team will be able to easily use to access their analyzed records and query words, and compare senses with senses.

The development language for the SSMS interface is DELPHI 7.0. Based on the need for program execution, the architecture of SSMS is shown in *Figure 6.* The SSMS has different functions, and these functions have been represented in different sub-windows of main-windows interface and with inter-linked ASP pages. Sense management and sense visualization are two major functions in SSMS. With the Sense management function, the Chinese Wordnet term can insert, update, and delete datas including lexical entries, word senses, meaning facets, POS, example sentences, English synset(s), and lexical semantic relations. The sense visualization function is the SSMS interface and can be divided into two parts: sense query and lexicon report. The main interface of SSMS is shown in *Figure 7.* The SSMS interface provides a user-friendly interface for operation and maintenance. With the sense query function, users can enter a serial number or a lexical entry for sense querying using the SSMS interface. *Figure 8* shows the system view of the query sense function. Various information about individual words can be shown in the working window, include synonyms, antonyms, hyponyms, and variants. Through the clear presentation, lexical semantic relations can be understood and compared. Another function, the lexicon report, uses the development software Crystal Report9 to produce electronic documents as shown as *Figure 9.*

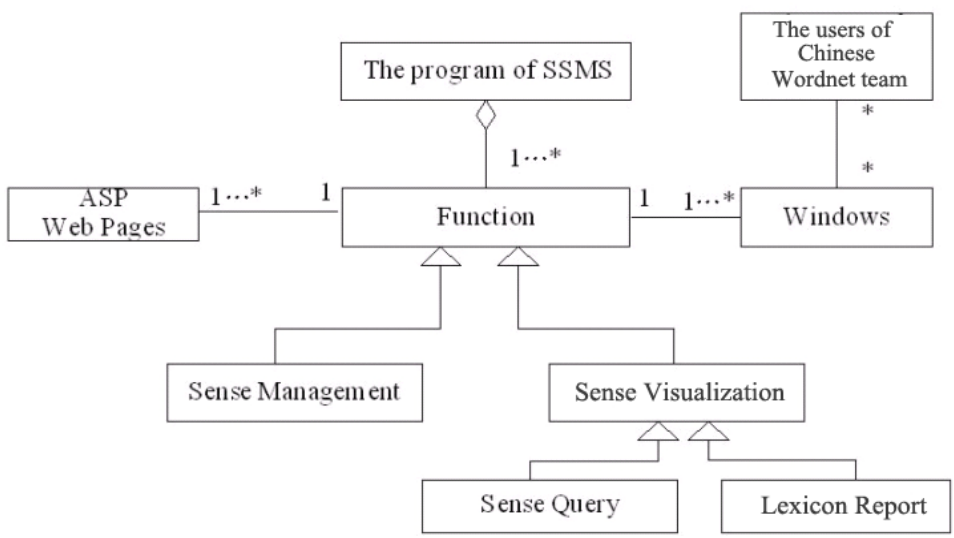

Figure 6. The class diagram of an SSMS function description

Figure 7. The main interface of SSMS

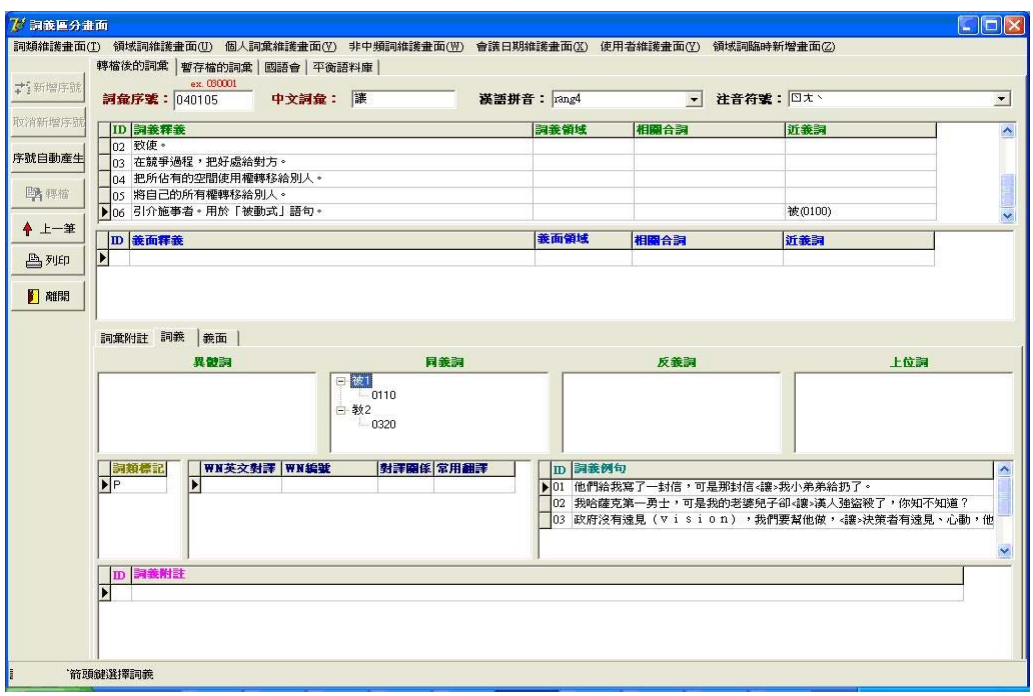

Figure 8. The system view of the query sense function

拌 ban4 ㄅㄢˋ
詞義 1：【及物動詞，VC】將任何材料混和、攪在一起。異體詞「伴」。{ blend, 00274169V }
義面 1：將食材和食材，或食材和醬汁混和在一起。
例句：小女孩剛開始是給小黃吃鮮魚**拌**飯，小黃吃得津津有味。
例句：植物油及動物油的含量如何能攝取到最低的程度，通常**拌**沙拉以植物油處理。
例句：他把牛肉切薄片，並舀一點湯燙一下就上桌，而乾麵則只加一些醬油**拌**一**拌**，吃不到師傅的手藝。

義面 2：將一般液態和固態的材料混和在一起，使成爲漿狀。
例句：他交給我一份以石灰水**拌**製海砂混凝土的研究計劃報告文件。
例句：牛屎伯公口裡從來沒停止過嚼檳榔，雙手則忙著**拌**紅灰、包荖葉。

附註：
1.分義面的主要原因在於「拌」這個詞用在食物上的用法頻率上相當高，已經形成特殊用法，所以以義面方式處理，以標誌出這種情形。

Figure 9. The format of a lexicon report

5. Results and Discussion

There is no question that semantic knowledge is needed to solve many problems in natural language processing. Building a sense-based lexical knowledgebase will be a subjective process, and the quality of this lexical knowledgebase will be highly dependent on the criteria which it must satisfy. In this paper, we introduced five criteria as well as operational guidelines for sense distinction and described how the Sinica Sense Management System can be used to manage both lexical entries and word senses.

More than 10 members of Chinese Wordnet Team were involved in the project. When we began this work, middle-frequency words were selected as the main terms to be analyzed. There are 2,018 middle-frequency words in Sinica Corpus. To date, 3,344 lemmas have been analyzed, and 5,914 senses have been identified. Among these records, the word “xia4” includes 42 senses and is the most complex term analyzed so far.

While more work needs to be done to improve the quality of this sense-based lexical knowledgebase, the goal of refining the procedure for word sense distinction needs to be accomplished at the same time. We believe that is a well-designed sense-based knowledgebase that it will serve as an important tool in Chinese knowledge processing applications.

References

Ahrens, K., L. Chang, K. Chen, and C. Huang, "Meaning Representation and Meaning Instantiation for Chinese Nominals," *Computational Linguistics and Chinese Lnaguage Processing*, 3, 1998, pp.45-60.

Booch, G., J. Rumbaugh, and I. Jacobson, *The Unified Modeling Language User Guide*, Addison-Wesley, 1999.

Fellbaum, C., et al., *WordNet: An Electronic Lexical Database*, MA: MIT Press, Cambridge, 1998.

Huang, C., et al., *Meaning and Sense in Chinese, 2nd Edition Technical Report 05-01*. CKIP, Academia Sinica, Taipei, 2005.

Huang, C., "Sense and Meaning Facet: Criteria and Operational Guidelines for Chinese Sense Distinction," Presented at *The Fourth Chinese Lexical Semantics Workshops*, 2003, Hong Kong City University, Hong Kong.

Muller, R.J., *Database Design for Smarties: Using UML for Data Modeling*, Morgan Kaufmann, 1999.

Oestereich, B., *Developing Software with UML Object-Oriented Analysis and Design in Practice*, Addison-Wesley, 1999.

Online Resources

Sinica BOW: http://BOW.sinica.edu.tw/

Sinica Corpus: http://www.sinica.edu.tw/SinicaCorpus/

WordNet: http://www.cogsci.princeton.edu/~wn/