The Sinica Sense Management System: Design and Implementation

Chu-Ren Huang, Chun-ling Chen, Cui-Xia Weng, and Keh-jiann Chen

Academia Sinica

1. Background and Motivation

It has been a trend for language engineering to construct a sense-based lexical knowledgebase as a core foundation. WordNet and Euro WordNet are two well-known examples. There are two important criteria in constructing this knowledgebase: linguistic felicity and data cohesion. Huang et al. (2003) discussed how to achieve linguistic felicity in building a comprehensive inventory of Chinese senses from corpus data. It introduced five criteria as well as operational guidelines for sense distinction. In this paper, we will discuss how to achieve data cohesion for the sense information thus collected through a Sinica Sense Management System (SSMS).

2. Introduction to the Content of the SSMS

The SSMS manages both lexical entries and word senses. This system is designed and implemented by the Chinese WordNet Team at Academia Sinica. It contains all the basic information that can be merged with the eventual Chinese WordNet. The basic structure of this system 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 sense and meaning facet, this system also includes the following information: POS, example sentences, corresponding English synset(s) from Princeton WordNet, and lexical semantic relation such as synonym/antonym, and hypernym/hyponym. Moreover, the overarching structure of the system is managed by a sense serial number, and inter-entry structure is established by cross-references among synsets and homographs.

In the present stage, the Chinese WordNet Team focuses on analyzing middle-frequent words in Sinica Corpus. The reason to choose middle-frequent words as our target ones is that with only three to five senses of a word, we can investigate senses and meaning facets of each word deeply and accurately, which would avoid the simple situation of one sense in low-frequent words, and the complicate situation in high-frequent words with numerous senses. Up to now, 1000 more lemma have been analyzed, and more than 2000 senses have been distinguished. We also published five technical reports to present these results [4]. In the near future, these fruits will be used as a basis for Natural Language Processing or E-learning application.

3. The Design Principle of SSMS

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

3.1 The Unique Serial Number

First, 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 any logical structure to it. Hence, although it guarantees unique identification, it is not very trackable. An alternative 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. And if decision is made to encode only certain higher level nodes, the random assignment issue is unavoidable since more than one lexical sense will be assigned to the same node. In our system, the unique serial number of each sense is composed of three segments: the sequential information of 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)" for example. "bao4 zhi3" has two senses and two meaning facets being distinguished. 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。

Four-level unique serial number is shown as below to express four segments of the unique serial number for one meaning of "bao4 zhi3".

報紙 "bao4 zhi3 (newspaper)"

Lemma processing year	03-			
Lemma form ID		-0018-		
The first sense			-01-	
The first meaning facet				-01

The unique serial number for 1st. meaning facet of 1st.sense of "bao4 zhi3" => **0300180101**

There are four advantages to manage the sense database with unique serial numbers. First, the sequential number not only gives a unique code to 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 facilitate man-machine interface such as in keyword search for senses. Third, it also provides a logical structure of 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 belong to each lemma. The first two digits are reserved for senses and the last for meaning facets. These four digits also allow the minimal space 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 contributes to trackability.

3.2 The Cross-reference device

Second, SSMS will automatically prompt all possible cross-references. When a lemma is called up for analysis, all existing records that contain this lemma will be prompted. This includes not only lexical semantic relations such as synonyms and hyponyms, it also includes

and sense definition that contain this lemma, as well as any explanatory notes that contain this lemma. This feature allows sense relations to be clearly defined, and inconsistencies to be detected. In addition any anomaly in definition or expression format will also be discovered. This process will also help us to narrow down to a set of control vocabulary for sense definition. This feature contributes to both the trackability of senses and consistency of sense definition.

3.3 The concurrent lexical knowledgebase and coupus

Third, SSMS enables parallel concurrent of the lexical knowledgebase and corpus. When a lemma is chosen in the system, all tagged example of that lemma from Sinica Corpus are retrieved. This allows closer examination of how the senses are used and distributed. It also allows automatic selection of corpus example sentences. In turn, when the sense classification is completed, SSMS allows all the corpus sentences to be sense-tagged and returned to merge with the original corpus. In other words, a sense-tagged corpus is being processed in parallel. This feature allows each lexical sense to be trackable to its actually uses in the corpus. It also allows linguist to examine the data supporting each sense classification.

3.4 Linking to the Sinica BOW

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

4. The Implementation of SSMS

There are three major phases in this system implementing. In lemma analysis phase, based on the criteria and operational guidelines proposed in Huang (2003), we distinguish senses and meaning facets for each word. At the same time, Sinica Corpus and WordNet will be referred for POS, examples and English translation. Then through the help of dictionary resources or word mapping by the system, we decide the word relation. The second phase can be divided into two steps. First, we design the schema of the sense management system database for storing the analyzing result of the first phase. Then, as for the data access, we develop the interface to help the Chinese Wordnet Team insert and query from the database. We employ DELPHI tool to design our system interface. Thought the interface, the data in the database also can be exported as Word documents. Last, the third phase of this system implementation is the application phase. Our work project is to build Chinese WordNet web sites for users querying. The development language of these web pages is HTML and ASP. Finally, these web pages in the web sites could be viewed thought web server. By the way of the Internet, people can 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.

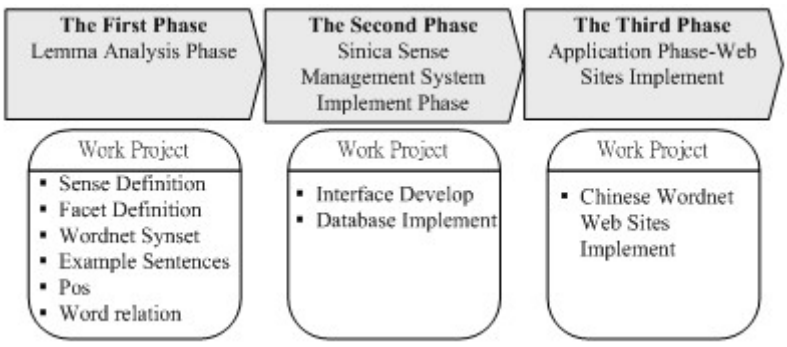


Figure 1: The flow chart of the Sinica Sense Management System.

We can represent the overall framework of SSMS diagrammatically in *Fig. 2*. As the diagram indicates, the Chinese WordNet Team use SSMS to access database and have electric documents as Word report. Moreover, the users in the internet can browse HTML/ASP pages to query database through and web server.

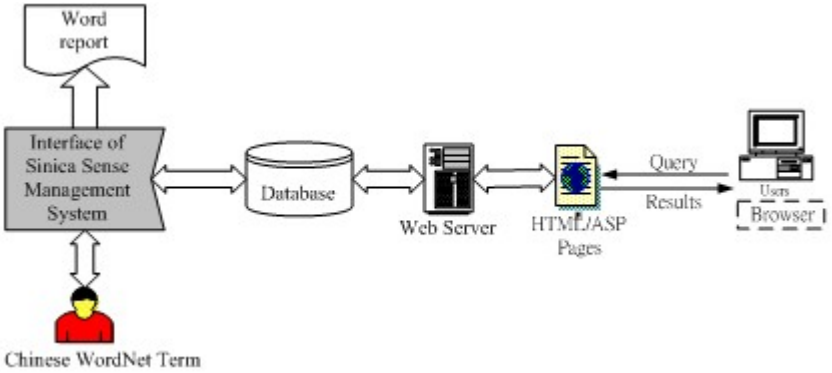


Figure 2: The overall structure of SSMS.

4.1 The Schema of SSMS Database in Class Diagram

In the section, we discuss and design the schema of SSMS Database. The Unified Modeling Language (UML) [2][7] is a graphical notation that provides the conceptual foundation for assembling a system out of components from the 4+1 views and nine diagrams. Each view is a projection into the organization and structure of the system, focused 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 only introduce the details about class diagrams as follows.

Class diagrams [2][7][6] 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 as shown in *Fig. 3*. The terms with *italic style* in *Fig. 2* indicates the concepts about class diagrams.

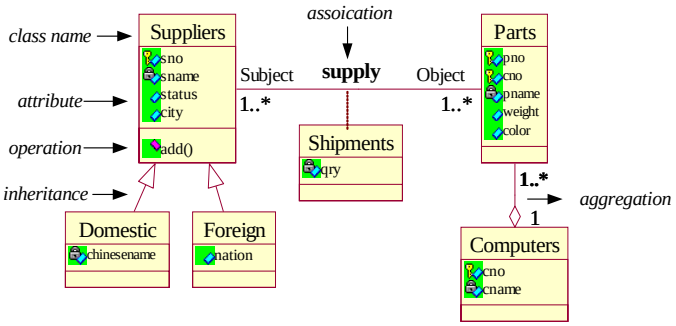


Figure 3: A class diagram for the Suppliers-and- Parts Database.

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 explain these terms as follows:

   (a) Class name: every class must have a name to distinguish it from other classes. For

example, **Suppliers** or **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 *Fig. 3*, the **Suppliers** have 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 to affect behavior. A class may have any number of operations or no operations at all. For example, in *Fig. 3*, the class of **Suppliers** has an 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 *Fig. 3*, 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 classes) and a special (subclasses) 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 *Fig. 3*, **Domestic** and **Foreign** suppliers (two subclasses) are a kind of suppliers (the super class).

According to the need of SSMS content and design principle, *Fig. 4* is the schema of SSMS database using the concepts of class diagram.

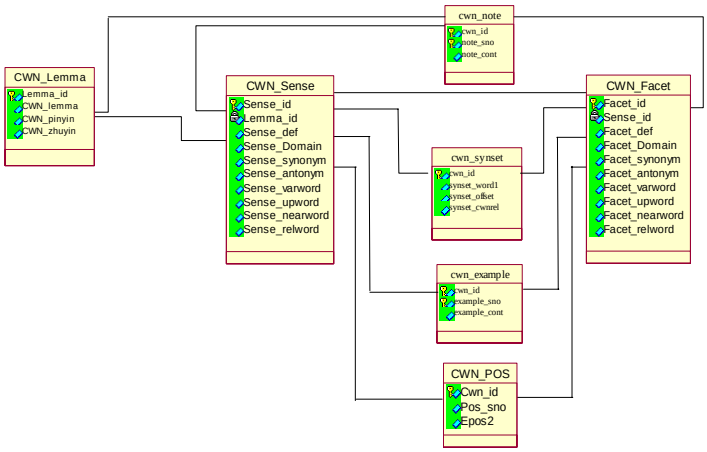


Figure 4: The schema of the Sinica Sense database.

4.2 The Function of SSMS

In this section, we will discuss the interface marking for SSMS. The development language of SSMS interface is DELPHI 7.0. Based on the need of program execution, the

function of SSMS is shown in *Fig. 5*. In SSMS, the programs have many functions and these functions can be represented in windows interface and ASP web pages. Sense management and Sense visualization are two major functions in SSMS. In Sense management function, the Chinese WordNet term can insert, update, and delete data including lexical entries, word sense, meaning facet, POS, example sentences, English synset(s), lexical semantic relation. The Sense visualization is SSMS interface and can be divided into two parts: Sense Query and Word Report. The format of SSMS interface is shown in *Fig. 6*. The SSMS interface provides a user-friendly interface to operate and maintain. For the Sense query function, the users can enter a serial number or a lexical entry for sense querying in SSMS interface. Another function, the Word report, uses development software Crystal Report9 to produce electric documents shown as *Fig. 7*.

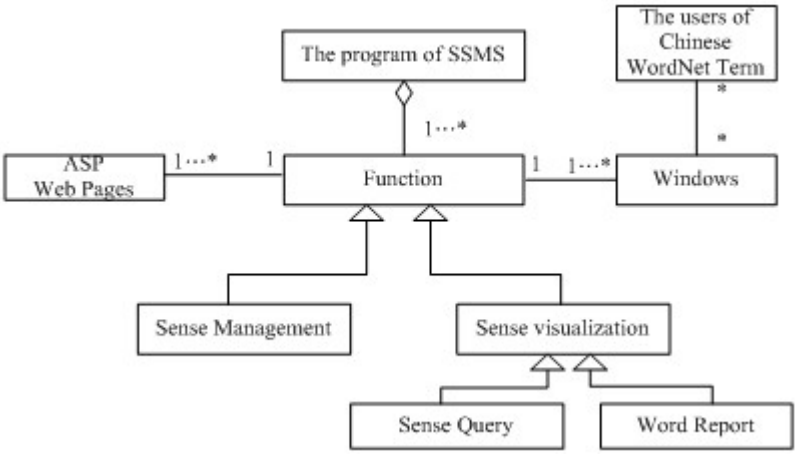


Figure 5: The class diagram of SSMS function description.

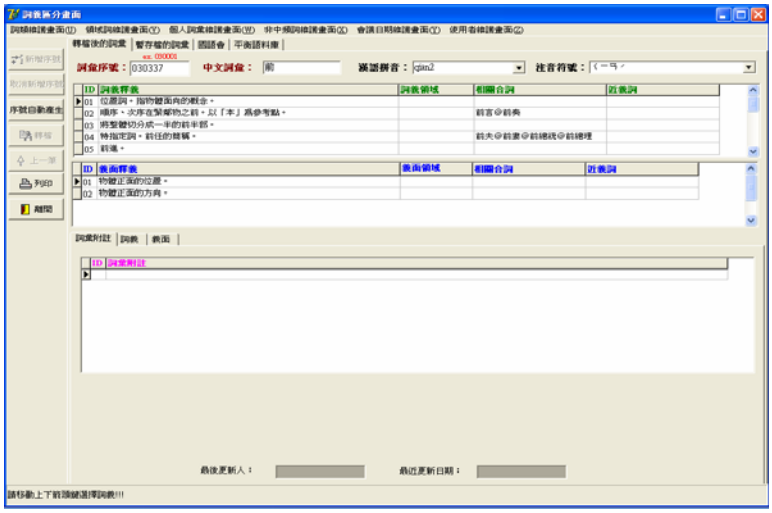
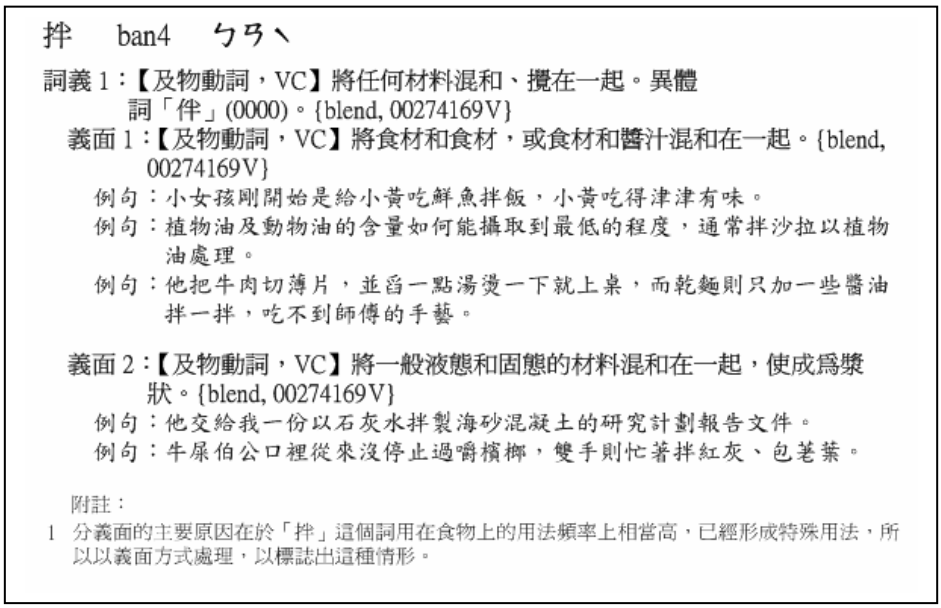


Figure 6: The interface of SSMS.



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」(0000)。{blend, 00274169V}

義面 1：【及物動詞，VC】將食材和食材，或食材和醬汁混和在一起。{blend, 00274169V}

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：【及物動詞，VC】將一般液態和固態的材料混和在一起，使成爲漿狀。{blend, 00274169V}

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 7: The format of Word report.

6. Conclusion

In sum, SSMS is not only a versatile development tool and management system for 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.

Online Resources:
Sinica BOW: http://BOW.sinica.edu.tw/
Sinica Corpus: http://www.sinica.edu.tw/SinicaCorpus/
WordNet: http://www.cogsci.princeton.edu/~wn/

References

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

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

[3] Fellbaum, Christine. Ed. 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[4] Huang, Chu-Ren (ed.), 2004, *Sense and Sensibility series: Technical Report 03-01~04.* CKIP, Taipei.

[5] Huang, Chu-Ren et al., 2003, Sense and Meaning Facet: Criteria and Operational Guidelines for Chinese Sense Distinction]. Presented at the Fourth Chinese Lexical Semantics Workshops. June 23-25 Hong Kong, Hong Kong City University.

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

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

CLSW5 Submission Information
Title: The Sinica Sense Management System: Design and Implementation

Authors: Chu-Ren Huang, Chun-ling Chen, Cui-Xia Weng, and Keh-jiann Chen

Affiliation: *Academia Sinica*
Contact Information:
churen@gate.sinica.edu.tw (Huang)
chunling@gate.sinica.edu.tw (CL Chen)
cxw@gate.sincia.edu.tw (Weng)
kchen@iis.sinica.edu.tw (KJ Chen)