# THE EVOLUTION OF MACHINE-TRACTABLE DICTIONARIES

*Cheng-ming Guo  Chang-ning Huang  Jun-ping Gong  Jin Li*

Computer Science Department
Tsinghua University, Beijing 100084 CHINA

## 1. Introduction

The usefulness of Machine-Tractable Dictionaries (MTDs) in automatic sense tagging of running text is clearly demonstrated in Tong, et al., (1993). While previous work (Yarowsky, 1992; Gale, et al., 1992, 1993) relies heavily on the role of statistics, Tong's system makes use of one MTD as well as two Machine-Readable Dictionaries (MRDs) in an example-based reasoning process in tagging sense numbers to novel words, compounds words, and phrases in the input text. The hit rate of correct sense tagging runs as high as 90%. The lowest hit rate ever recorded was 70%. The system is considered a necessary mechanism for the construction of annotated Chinese *Monitor Corpora* (Sinclare, 1991) from running Chinese text.

Earlier work (Wilks, et al., 1990) identified a distinction between MRDs and MTDs. This paper recognizes a further distinction between two types of MTDs, i.e., Type I and Type II MTDs. Type I MTDs are built on the definition primitives of one particular MRD whereas Type II MTDs are built on the semantic primitives of one particular natural language. The MTD functioning in Tong's system belongs to the Type I category. The evolution from Type I to Type II MTDs represents an evolution of the primitives the MTDs are constructed on. The authors believe that, although it is feasible to derive a natural set of definition primitives from one particular MRD, a descriptive set of semantic primitives (Wilks, 1977) of one particular natural language is preferably derived from more than one natural source such as a dictionary. This position represents a backoff from the previous one that claims the derivation of a natural set of semantic primitives from one particular MRD, particularly *the Longman Dictionary of Contemporary English* (LDOCE) (Guo, 1989).

Three Type I MTDs completed at Tsinghua University will be presented. These include MINI-LDOCE constructed for Longman Dictioanries, the MTD version of *The Modern Chinese Dictionary of General Chinese Characters* (MTD-XIANTONG) (Fu, 1987), and the MTD version of *The Multifunctional Dictionary of Modern Chinese Words* (MTD-DUOGONGNENG) (Feng & Zhou, 1989).

Efforts to derive natural sets of semantic primitives from Type I MTDs are then described. These include work on detecting circular defintions within one particular MTD, work on computing the compatibility of word senses between two monolingual dictionaries, and the most recent step at deriving a natural set of Chinese semantic primitives from Type I MTDs of Chinese. The work was done on Sun-4 workstations with the C computer language.

## 2. Design and Construction of Type I MTDs

The design and construction of MTDs in general are discussed in detail by various authors in Guo (1994). Described in this paper are three Type I MTDs in two languages. The first one is the English MINI-LDOCE designed and constructed by Cheng-ming Guo for Longmans, the second the MTD-XIANTONG constructed by machine by Xiang Tong (Tong, et al., 1993), and the third the MTD-DUOGONGNENG by Jun-ping Gong (Gong & Huang, 1991) by hand.

### 2.1. MINI-LDOCE

The discussion presented below represents a general description of the most essential points in the design and construction of MINI-LDOCE. For a detailed account of the project the reader is referred to the MINI-LDOCE system documentation upon Longmans' publication.

As is well known, Longman Dictionary of Contemporary English (LDOCE) has a Controlled Vocabulary (CV) of about 2,000 words. These 2,000 CV words are used in every word sense definition of the entire LDOCE. It was found that a subset of about 4,000 CV word senses are used to define the 2,000 CV words themselves (Guo, 1989). MINI-LDOCE is a dictionary of these 4,000 definition primitives of LDOCE. In other words, MINI-LDOCE has about 4,000 entries of word senses, each entry of a word sense being a definition primitive of LDOCE.

MINI-LDOCE differs from LDOCE in the form of the definition text. Whereas LDOCE word senses are defined by words, MINI-LDOCE definitions are given in word senses, i.e., each word in the definition text is already disambiguated, and has a sense number attatched at the end. The set of the CV word senses used in the definition text of MINI-LDOCE falls within the set of total MINI-LDOCE sense entries. When non-CV words, or phrases formed either of CV words, non-CV words, or a combination of CV and non-CV words are found in MINI-LDOCE sense definitions, an effort is made to disambiguate the non-CV word or phrase, and reduce its definition as found in LDOCE to the definition primitives of LDOCE. This results in embedded definition, which sometimes runs down three or four levels. In most cases, the embedded definition *bottoms out* to definition primitives within three levels of embedded definitions.

One important motive behind the design and construction of MINI-LDOCE is its expected role as an important, although apparently meagre, source of lexical and knowledge data in a process called *bootstrapping*, i.e., to convert, in steps, the rest of LDOCE into a Type I MTD of the MINI-LDOCE kind.

MINI-LDOCE has been completed, and is now available to both academic and commercial users. Besides its possible use as the nucleus of English lexicon, its construction constitutes a necessary first step in generating a natural set of semantic primitives for the English language. As mentioned above, the semantic primitives form the basic units for the construction of Type II MTDs in the same way definition primitives form the basic units for that of Type I MTDs.

## 2.2. MTD-XIANTONG

MTD-XIANTONG is the MTD version of *The Modern Chinese Dictionary of General Chinese Characters* (Fu, 1987). XIANTONG is the shortened name for the original Chinese dictionary. It contains about 10,000 monosyllabic Chinese words in the form of characters with more than 50,000 multisyllabic Chinese words and phrases as examples. The meanings of every monosyllabic Chinese word are defined in the sense definitions given under the entry word in question. However, the multisyllabic words and phrases given as examples following each word sense definition do not have accompanying definitions. Each word sense definition is separated from its examples by a colon. An example is given below:

康 1。健康：康复 | 康健 | 康宁 | 安康
2。广大：康衢 | 康庄 | 康庄大道
3。安，乐：康乐 | 小康
4。姓

The construction of MTD-XIANTONG proceeds in two steps. At Step 1, more than 50,000 multisyllabic example words and phrases are automatically sense tagged. At Step 2 the definition text of every word sense of all head words are disambiguated, attaching a sense number to every word in the sense definitions. Note that the MTD version of XIANTONG contains entries of words in frequent use only. Monosyllabic Chinese words that are not in frequent use (2,193 of them) stay out of MTD-XIANTONG. Word sense entries in MTD-XIANTONG total around 15,000 with about 50,000 examples of compound words and phrases. All words in the definition texts and all examples are tagged with word sense numbers.

The algorithm used in automatically sense tagging the multisyllabic words in XIANTONG is fairly simple. It takes advantage of one important fact concerning the making of the original Chinese dictionary, i.e., repeated appearance of the same example word and phrase under different head words, or more exactly, under different sense definitions of different head words. In fact, all head words in XIANTONG are single Chinese characters. For example a bisyllabic Chinese word made up of two Chinese characters may appear twice as an example under the two composite characters, each time as an example of a particular word sense of one of the two characters being defined. For example, the word 康复 (recover) appears under both 康 and 复, 康复 appears under 康 as an example in Sense 1 (p.275), and under 复 as an example in Sense 1 (p.151), too. Hence 康_A01 复_A01. By checking through all example words and phrases in the way they are cross-referred, all words that are cross-referred as examples in the dictionary are automatically tagged with appropriate sense numbers. This process produced the bulk of MTD entries which form the core of the reasoning system in the automatic sense tagging of running Chinese text (Tong, et al., 1993).

At Step 2 the automatic tagging of sense numbers to each word in the definition text proceeds in the same way as described in Tong, et al. (1993).
Efforts are now being made to blend in human judgement to the making of a more accurate version of MTD-XIANTONG.

## 2.3. MTD-DUOGONGNENG

MTD-DUOGONGNENG is the MTD version of *The Multifunctional Dictionary of Modern Chinese Words* (Feng & Zhou, 1989). DUOGONGNENG is the shortened name for the original Chinese dictionary. It contains about 7,000 most frequently used words in everyday Chinese.

MTD-DUOGONGNENG is produced by manually tagging word sense numbers to each word in the definition text of every sense entry under every head word of the entire dictionary. Unlike MTD-XIANTONG which uses its own definition primitives, MTD-DUOGONGNENG uses the definition primitive of another Type I MTD, i.e., MTD-XIANTONG. Hence the tagged sense numbers in MTD-DUOGONGNENG agree with the system of sense numbers of MTD-XIANTONG. This paves the way for the generation of a set of descriptive semantic primitives from more than one MTDs.

In the corpus of the definition texts of MTD-DUOGONGNENG, 10,003 word types appear with 5,873 definition primitives. The 5,873 definition primitives are the word senses of 2,574 monosyllabic Chinese words in the form of Chinese characters from MTD-XIANTONG. The total number of word senses given in MTD-XIANTONG of these 2,574 Chinese characters amounts to 9,363.

## 3. Evolution from Type I to Type II MTDs

The absolute advantage of Type II MTDs over Type I MTDs is an empirical question. However, the advantage of a descriptive set of semantic primitives over a set of definition primitives derived from one single data source should be obvious since the former covers all possible source data of one natural language.

Using a Type II MTD on parallel machines, especially when a descriptive set of semantic primitives are used as microfeatures of an Artificial Neural Network (ANN) system constitutes an interesting research to both ANN and MTDs. Further still, Type II MTDs may help revive the dated notion of turning symbolic processing of natural languages into mathematical computation by using a function that turns Type II MTD definitions based on semantic primitives into mathematical representations.

### 3.1. From Conceptual Equivalents to Circular Definitions

Circular definitions may exist within definitions primitives derived from a dictionary. For instance, both *bottom1* and *base1* may fall into a set of definition primitives derived from LDOCE. However, they are locked in a definition circle:

*bottom1:* the base on which something stands; the lowest part, inside or outside
*base1:* the bottom of something

Desirably, a descriptive set of semantic primitives is free of such circular definitions between members of the set. Karen Sparck-Jones(1967) conducted an in-depth study on circular definitions in English dictionaries. Similar study on Chinese dictionaries did not come true until recently when Chinese MRDs are available. Below are some results from a similar study on MTD-XIANTONG.

The basic idea is for *Circular Definitions* to emerge out of sets of *Conceptual Equivalents*. The so-called *Concerptual Equivalents* refer to a set of related conceptual entities derived from the sense definitions of a MTD. The so-called *Circular Definitions* refer to

definition circles found in a set of *Conceptual Equivalents*. A study which takes monosyllabic-word concepts as search units reveals that MTD-XIANTONG has 389 sets of conceptual equivalents among 15,273 sense definitions. The 389 sets of conceptual equivalents contain 1,716 sense definitions which amount to 11.23% of the total. 28 strict circular definitions are spotted among the 389 sets of conceptual equivalents. See the following for an example:

a. conceptual equivalents:

反_A02　　返_A01　　复_A04

归_A01　　还_B02　　回_A01

b. circular definition

回_A01 <--- 返_A01 <--- 回_A01

### 3.2. Sense Compatibility Between Monolingual Chinese Dictionaries

Sense compatibility computation was studies in Byrd, et al. (1987) as a mapping problem between dictionaries. The mapping was seen as a *symmetric binary relation* between two different dictionaries. In the present study on the derivation of a descriptive set of semantic primitives, sense compatibility computation is seen as an important element in merging data from multiple dictionary sources.

Two dictionaries XIANTONG and *New Modern Chinese Dictionary* (henceforth XIANHAN) (Wan, et al., 1985) are used as data. Two sense definitions under the same head word in the two dictionaries are compatible if:

**a**. the two definition texts are identical;

**b**. one definition text is included as part of the other;

**c**. all segments of one definition text separated by punctuations appear in the other definition text;

**d**. more than half of the example words and phrases given after one definition text appear as examples after the other;

**e**. both have long definition texts, and the first ten words are identical.

Results show a total of 58.8% of all definition texts are compatible with detailed split as follows:

**a**. 23%　　**b**. 12.5%　　**c**. 10.5%

**d**. 6.8%　　**e**. 6%

On the basis of such findings, XIANTONG and XIANHAN are merged into one dictionary with 75,572 word sense definitions.

### 3.3. From Definition Primitives to Semantic primitives

The set of definition primitives of a Type 1 MTD close on the MTD. A descriptive set of semantic primitives close on all data of one natural language. The distinction between *descriptive* semantic primitives and *prescriptive* semantic primitives is given in Wilks (1977). Essentially, unlike prescriptive semantic primitives which are *prescribed*, so to speak, by man, descriptive semantic primitives are *derived* from natural sources such as a dictionary. This paper represents a backoff from Guo (1989) where it was claimed that a natural set of semantic primitives had been derived from LDOCE. It is now believed that although it is feasible to derive a set of definition primitives from one particular MRD or MTD, a *descriptive* set of semantic primitives true of one natural language is preferably derived from more than one natural source such as a dictionary.

What is presented below represents a first step to the generation of a set of Chinese semantic primitives from a set of definition primitives derived from one particular Chinese dictionary, i.e., MTD-DUOGONGNENG. A hunch set of semantic primitives, which is a subset of the definition primitives derived from MTD-DUOGONGNENG, is applied to the sense definitions of MTD-XIANTONG in an effort to find out how many word sense entries can be *accounted for*, i.e., defined, within five defining cycles.

As given above the definition primitives found in MTD-DUOGONGNENG total 5,873. 5,796 of them are senses of words that are known as most frequently used Chinese words. 4,417 out of these 5,796 definition primitives appear more than once in MTD-DUOGONGNENG sense definitions. These 4,417 definition primitives are chosen as the initial hunch set of semantic primitives. On the other hand, MTD-XIANTONG was chosen as the first other dictionary to try out these 4,417 defining primitives from MTD-DUOGONGNENG. A subset of 15,273 sense definitions of MTD-XIANTONG participate in the generation processs. This subset, 7,617 strong, belong to the most frequently-used words in contemporary Chinese. At the first defining cycle, 3,804 word senses are defined by the 4,417 hunch set of semantic primtives. Note that every time a word sense is defined, it joins the hunch set immediately to define more word senses in the next defining cycle. At the second defining cycle, an additional 556 word senses are defined. At the third defining cycle, another 144 word senses are defined. At the fourth defining cycle, 27 more word senses are defined. At the last defining cycle 4 additional word senses are defined. At the end of the fifth defining cycle, 3,256 word senses remain undefined. A complexity of reasons contribute to the failure. One major one among them is due to incorrect and incomplete data. As discussed above, MTD-XIANTONG was produced by machine. Words tagged with 000 or MIS indicate that the system was unable to tag the correct sense number either because the word sense is not found in the dictionary or because the system cannot come up with the correct sense nubmer during the reasoning process. Efforts are being made to verify the data exhaustively by human judgement.

### References


Feng, Z-C., and Zhou, X-J, (1989), *The Multifunctional Dictionary of Modern Chinese*. International Culture Publishing Corporation: Beijing.

Fu, X-L., (1987), *Modern Chinese Dictionary of General Chinese Characters*. Foreign Language Teaching and Research Publishing House: Beijing.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). Work on statistical methods for word sense disambiguation. *Working Notes for AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pp. 54-60.

Gale, W. A., Church, K. W., and Yarowsky, D. (1993). A method for disambiguating word senses in a large corpus. *Computers and Humanities*.

Gong, J-P., and Huang, C-N. (1991). A project of statistics & analysis of Chinese morphemes and senses. Proceedings of NLPRS-91, pp. 108-112.

Guo, C-M. (1989), *Constructing a Machine Tractable Dictionary from Longman Dictionary of Contemporary English.* PhD desertation, New Mexico State University.

Guo, C-M. (Ed.). (1994), *Machine-Tractable Dictionaries: Design and Construction.* Ablex: Norwood, NJ.

Sinclare, J. (1991), Monitor corpora. *Corpus, Concordance, Collocation.* Oxford University Press. pp. 24-26.

Sparck-Jones, K. (1967) *Dictionary Circles.* System Development Corporation.

Tong, X., Huang, C-N., and Guo, C-M. (1993), Example-based sense tagging of running Chinese text. Paper read at the First International Workshop on Very Large Corpora. Columbus, OH.

Wan, Q-Z., Min, F-L., Yu, X-M., and Wu, T-Z. (1985), *New Modern Chinese Dictionary.* Xinhua Publishing House: Beijing.

Wilks, Y. A. (1977), Good and bad arguments about semantic primitives. *Communication and Cognition,* 10, pp. 182-221.

Wilks, Y. A., Fass, D. C., Guo, C-M., McDonald, J. E., Plate, T., and Brian, B. M. (1990), Providing machine-tractable dictionary tools. *Machine Translation,* 5, pp. 99-154.

Yarowsky, D. (1992), Word sense disambiguation using statistical models of Roget's categories trained on large corpora. *COLING-92.*