[From: Makoto Nagao (ed.) *Machine translation summit*. Tokyo: Ohmsha, 1989]

3 • 8 Electronic Dictionary Project

Hiroshi Uchida, Takahiro Kakizaki
Japan Electronic Dictionary Research Institute Ltd., Japan

[1] Introduction

As the range of applications for computers continues to increase, so does the need for a technology enabling computers to process and understand the languages people use, natural languages. This technology, natural language processing, will be fundamental in expanding the applications of computers because it underlies the man-machine interface.

Some important applications of natural language processing are:

— Advanced document processing systems such as machine translation, information retrieval, automatic summarizing

— Advanced natural language interfaces such as question-answering systems

— Advanced CAI systems using natural language

The Japan Electronic Dictionary Research Institute (EDR) was established in April 1986 by the Japan Key Technology Center and eight private companies (Fujitsu, Ltd., NEC Corporation, Hitachi, Ltd., Sharp Corporation, Toshiba Corporation, Oki Electric Industry Co., Ltd., Mitsubishi Electric Corporation, and Matsushita Electric Industrial Co., Ltd.,) to develop an electronic dictionary system for natural language processing. The purpose of EDR is not only to develop an electronic dictionary, but also to make it a standard.

[2] Electronic Dictionaries

At the present, there are a number of "electronic dictionaries" available which duplicate the contents of published dictionaries in storage media, and have procedures for accessing those contents. However, these electronic dictionaries are meant for humans to use and understand, and differ from the electronic dictionaries we will discuss here.

The electronic dictionary is not simply a machine-readable dictionary; it is a dictionary containing all the information necessary for computers to understand natural language.

Both people and computers must know the meaning of word to understand natural language text, The meaning of words, i.e. the concepts expressed by words, the grammatical characteristics of words when they express concepts, and knowledge necessary for understanding concepts, is contained in the electronic dictionary.

Large-scale electronic dictionaries are being developed or are already available for machine translation systems. Small-scale electronic dictionaries are being developed for question answering, discourse understanding, and speech recognition systems. However, each dictionary contains only the information necessary for the application in question, and the description format is idiosyncratic. The reason for this is that the contents reflect the grammar rules, algorithms, etc. of the system.

EDR is developing electronic dictionaries which are not restricted to use in particular applications. Information relating to grammar rules or algorithms is

excluded; only information about words and the concepts they express will be recorded. Following completion, the dictionary will be evaluated qualitatively and quantitatively by using it in systems for machine translation, speech recognition, etc.

[3] Dictionary Structure

The dictionary is composed of word entries and concept entries. Word entries contain the following information:

(a) a headword, the surface expression of a concept

(b) concepts the headword expresses

(c) grammatical characteristics of the word when it expresses a concept

Explanations of concepts that a word expresses are written as sentences, for people to distinguish one concept from another. These explanatory sentences are called concept heads, and are used as the head of concept entries.

Concept entries describe the relations that hold between two concepts. Compound concepts are described using more fundamental concepts. They are defined by the collection of binary relations between the concepts. Relations between concepts include case relations, causal relations, and synonym, similarity, superordinate and subordinate relations. Fig. 3-6 shows the relationship between word and concept entries.

In Fig. 3-7, the word "eagle" has more than two concepts, including "a score of two below on any hole", as used in golf, and "a bird called eagle." The grammatical characteristic for both is noun. The word "fly" also has more than one concepts, including "to move through the air on wings", with the grammatical characteristic intransitive verb, and "an insect called fly", with the grammatical characteristic noun.

The concept dictionary defines probable relations between these concepts, so the concept "a score of two below on any hole" can express the "degree" of the concept "to reach the end of an activity." The concept "a bird called eagle" can be an "agent" of "to move through the air on wings."

As Fig. 3-8 shows, ten types of dictionaries can be constructed from word and concept entries. There are two types of word dictionaries, the basic word dictionary (200,000 words) and the terminology dictionary (100,000 words). Eight dictionaries will result from constructing basic word dictionaries and terminology dictionaries for Japanese and English and the language pairs.

There are two types of concept dictionaries: a concept classification that describes only superordinate and subordinate relations, and a concept description that defines the other relations.

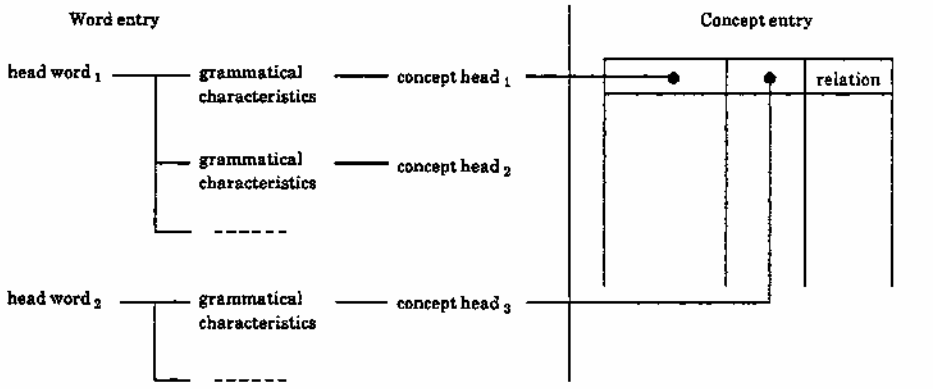

Fig. 3-6 Composition of electronic dictionary

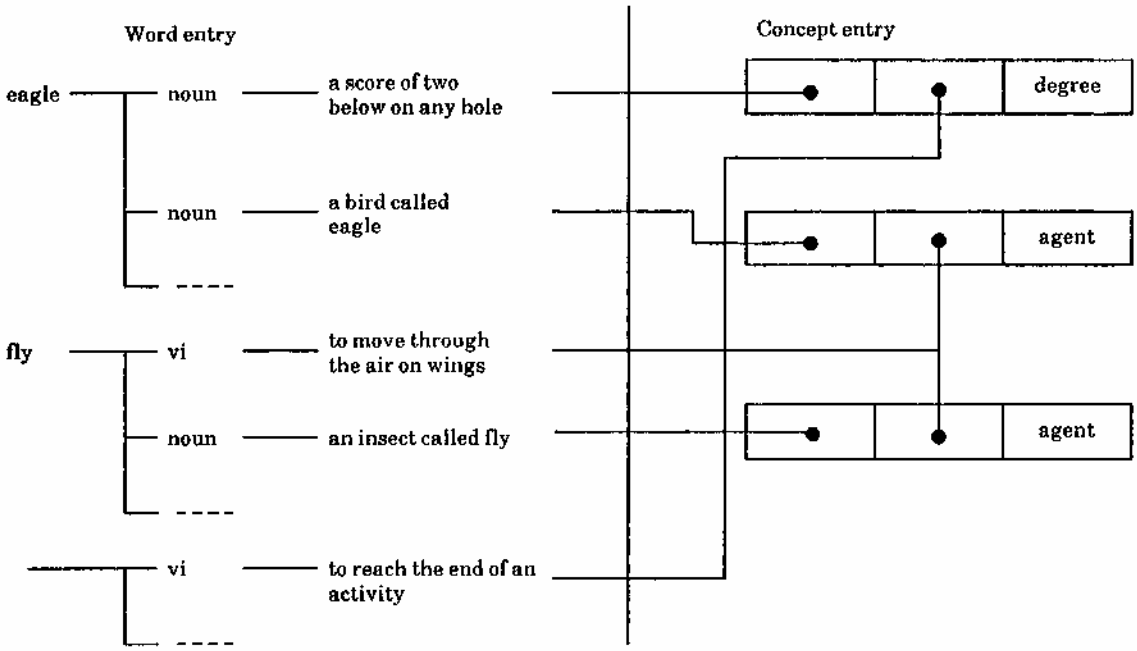

Fig. 3·7 An example of electronic dictionary

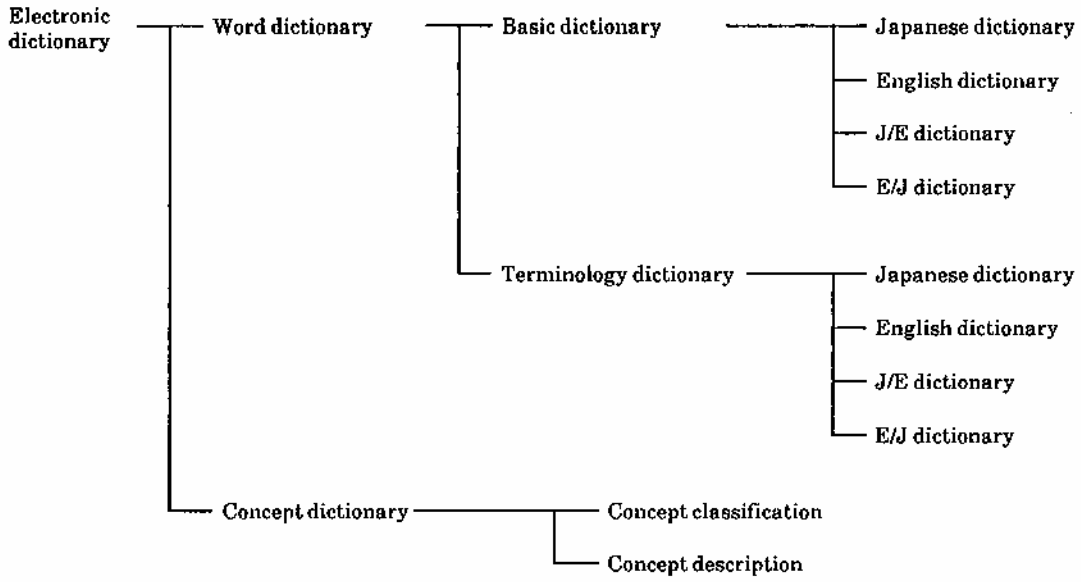

Fig. 3-8 Types of electronic dictionaries

[4] Production Method and Development Plan

The biggest problem in the development of a large-scale electronic dictionary is guaranteeing the uniformity and accuracy of information. At EDR, the worksheet shown in Fig. 3-9 is used in the production of the dictionary. A worksheet is created for each concept a word expresses, and data entry is performed later. Accuracy of the information is verified by analyzing a large-scale text base.

The concept dictionary is produced using the results of analysis of a large-scale text base. We will also develop machine translation systems, speech understanding systems, and information retrieval systems for testing and evaluating the dictionary. The results of the evaluation will be fed back to dictionary production. Fig. 3-10 shows the plan for dictionary development.

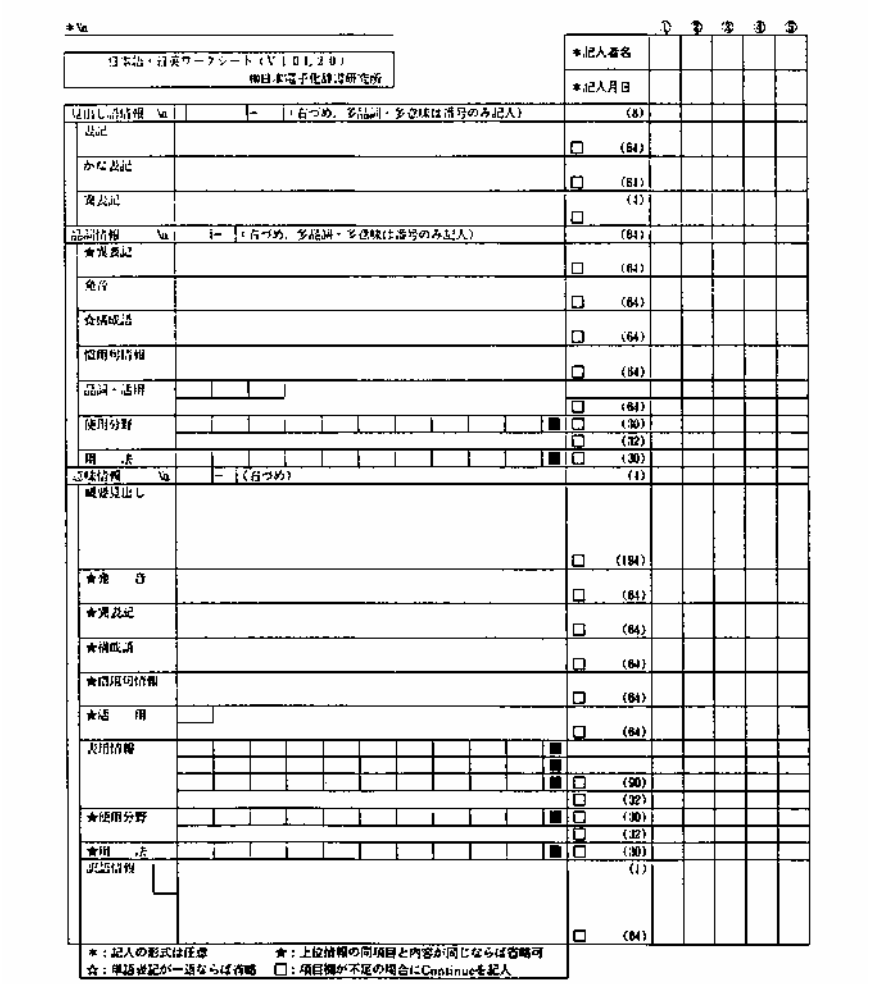

Fig. 3-9 Worksheet

[5] Usage of the Dictionary

The electronic dictionaries under development at EDR contain information necessary for computers to understand natural language. Applications include machine translation, question answering, and information retrieval. In particular, a machine translation system using an interlingua is considered the primary goal.

Morphological analysis and syntactic analysis, both a part of text analysis, are performed using grammatical characteristics in the word entries. Ambiguities arising in analysis are resolved by using the information about concepts in the concept descriptions.

[6] Conclusion

EDR is developing electronic dictionaries for Japanese and English. Electronic dictionaries for other languages must also be developed to support a multi-lingual machine translation system using an interlingua. In machine translation systems using the EDR dictionaries, links between languages (translation information) are established using synonym, similarity, superordinate, and subordinate relations, etc. described in the concept dictionary. It is our hope that dictionaries developed from now on will retain this concept dictionary based interface, widening the link between languages as dictionaries are produced.

Sub-theme \ Year	1986 fiscal year	1987 fiscal year	1988 fiscal year	1989 fiscal year	1990 fiscal year	1991 fiscal year	1992 fiscal year	1993 fiscal year	1994 fiscal year
General dictionary J, E, J/E, E/J)	Prototype 22,000 words	Prototype 100,000 words	Prototype	Evaluation	Improvement /Expansion	Improvement /Expansion	Improvement /Expansion-Evaluation	(Goal) 200,000 words for each dictionary	
Terminology dictionary J, E, J/E, E/J)		Prototype	Prototype	Prototype	Evaluation	Improvement /Expansion	Improvement /Expansion-Evaluation	(Goal) 100,000 words for each dictionary	
Concept classification	Experiment /Design	Primary prototype	Primary prototype	Evaluation	Secondary prototype	Improvement /Expansion	Improvement /Expansion-Evaluation		
Concept description	Experiment /Design	Primary prototype	Evaluation	Secondary prototype	Secondary prototype	Secondary prototype	Evaluation	Improvement /Expansion-Evaluation	Improvement /Expansion-Evaluation
Data management system	Design/ Prototype	Improvement /Expansion	Design/ Prototype	Prototype	Improvement /Expansion	Improvement /Expansion	Improvement /Expansion		
Testing and evaluation system		Design/ Prototype	Primary prototype	Evaluation /prototype	Secondary prototype	Secondary prototype	Actual use /Evaluation	Actual use /Evaluation	Actual use /Evaluation

Fig. 3-10 Research and development schedule