[MT Summit IV, July 20-22, 1993, Kobe, Japan]

The Electronic Dictionary Project — Current Status and Future Plans

Seibi Chiba
Japan Electronic Dictionary Research Institute, Ltd. (EDR)
Mita-Kokusai Building Annex, 4-28, Mita 1-chome, Minato-ku, Tokyo 108
phone: (03)3798-5521, fax: (03)3798-5335, e-mail: chiba@edr.co.jp

1. Introduction

The development of a large-scale electronic dictionary for natural language processing was planned as a 9-year national project by the Ministry of International Trade and Industry (MITI). The Electronic Dictionary Project began in 1986, and to carry the project into execution, the Japan Electronic Dictionary Research Institute, Ltd. (EDR) was established in April 1986 by joint funding from the Japan Key Technology Center, which is an organization founded by the government for promoting key technologies, and eight private corporations[1].

The objective of the Electronic Dictionary (ED) Project was to develop a large-scale dictionary that is optimized for and can be used effectively in computer processing. The dictionary is to be referred to as "EDR Electronic Dictionary", or for short, "EDR Dictionary". Since dictionaries of this kind are very important as common infrastructures indispensable for building the next generation of natural language processing systems and knowledge information processing systems, and since the scale of the dictionary is too large to compile individually, the eight major computer system manufacturers concerned joined in the project as funding members. The total fund of the project is expected to amount to about ¥14 billion.

In the background of the project, there existed critical needs, which resulted largely from the characteristics of the Japanese language, for electronic dictionaries or for natural language processing technology that requires electronic dictionaries. The needs are summarized as follows:

(1) Language barrier:

A great language barrier exists between Japan and other countries. The critical demand for natural language processing technology by computer, especially for machine translation technology, is a direct consequence of this barrier.

The Japanese language, moreover, has many unique characteristics. Accordingly, machine translation technology development has been seriously carried out at great expense. This development is a source of the critical need for electronic dictionaries and makes their specifications concrete.

2) Technological gap:

The processing technology of Japanese language is much inferior to that of European languages. To raise this technology level for documentation technology, language tutoring technology, and so forth, the implementation of computerized support systems is strongly desired.

(3) Core technology:

Natural language processing technology is becoming the core of information processing technology. In advanced knowledge and information processing and in artificial intelligence, natural language processing technology is not an application technology but a basic common technology that is needed in many fields.

2. Overview of the ED Project

The project aims at the R&D of a large-scale, high-level electronic dictionary (in other words, machine-tractable dictionary) necessary for the future generation of natural language processing and knowledge information processing technologies. The goal of the project, specifically, is to develop a general dictionary specification, a development methodology, and support systems without being affected by languages and applications. This goal also includes the cultivation of international and inter-industrial cooperation.

The goal is pursued by the efforts to achieve the following three sub-goals:

(1) To develop a dictionary that can be easily processed and recompiled with computers into various forms for specific purposes.

(2) To develop a dictionary by utilizing to a full scale the current computer and natural language processing technology.

(3) To develop a dictionary for computers to process and to understand languages.

To realize the sub-goals, six research themes were set up in the project. The overall schedule is outlined in Figure 1.

The EDR Electronic Dictionary encompasses both Japanese and English. The Word Dictionary describes mainly surface-level or grammatical information, and is divided into two categories: general vocabulary and technical term dictionaries. The Concept Dictionary describes deep level or semantic information that is suitably organized for handling by a computer, and is also divided into two categories: the Concept Classification and the Concept Description.

The Dictionary Management System and the Evaluation System are software systems. The former consists of writer's tools and user's tools. The latter is used for evaluation and improvement of the dictionaries.

The project adopted a distributed laboratory system with eight research laboratories. A research laboratory is located in each of the eight funding companies, and is connected to one another by a computer network for close collaboration. Most of the research staff are transferred from the funding companies. Every laboratory keeps in close contact with groups within the company that are developing their own large-scale application systems of natural language processing including machine translation.

It is also considered important to obtain assistance from outside research groups or institutions, which are supposed to become major users of the EDR Electronic Dictionary. EDR has therefore

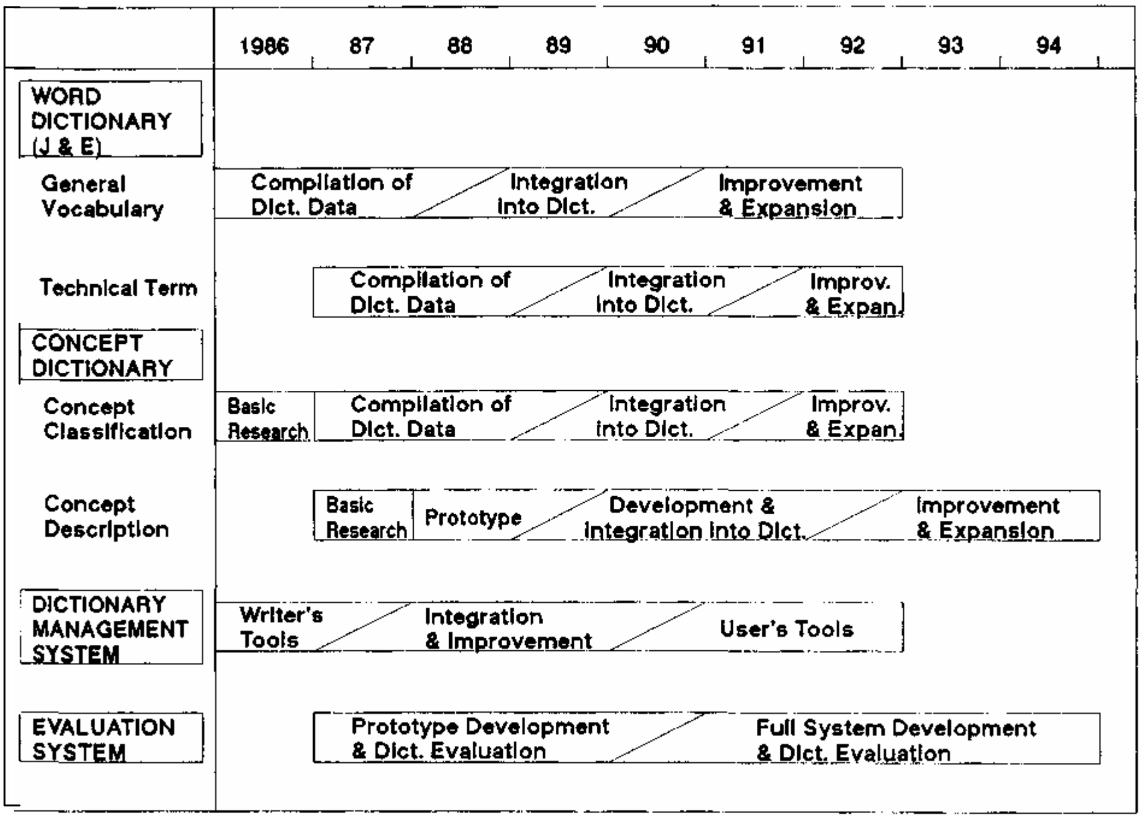

Figure 1. Schedule of the Electronic Dictionary Project

been promoting joint research programs with the Electrotechnical Laboratory (ETL), Institute for New Generation Computer Technology (ICOT), Center of the International Cooperation for Computerization (CICC) and several major universities in Japan including Kyoto University.

3. The Status of the EDR Electronic Dictionary Development

The EDR Electronic Dictionary is composed of four types of dictionaries (that is, Word Dictionary, Concept Dictionary, Cooccurrence Dictionary and Bilingual Dictionary), and the EDR Corpus as shown in Figure 2.

The Word Dictionary is divided into the General Vocabulary Dictionary and the Technical Term Dictionary. The General Vocabulary Dictionary is further subdivided into Japanese and English dictionaries of 200,000 words each. The Technical Term Dictionary covers the field of information processing, and also is split into Japanese and English dictionaries, each containing 100,000 words. The Word Dictionary contains the grammatical information and concept identifiers with concept illustrations. The concept in the EDR Electronic Dictionary represents a meaning of the word, and is defined independent of languages such as Japanese and English. The concept identifier is used as the interface or index for computers to refer to the Concept Dictionary. The concept illustration is

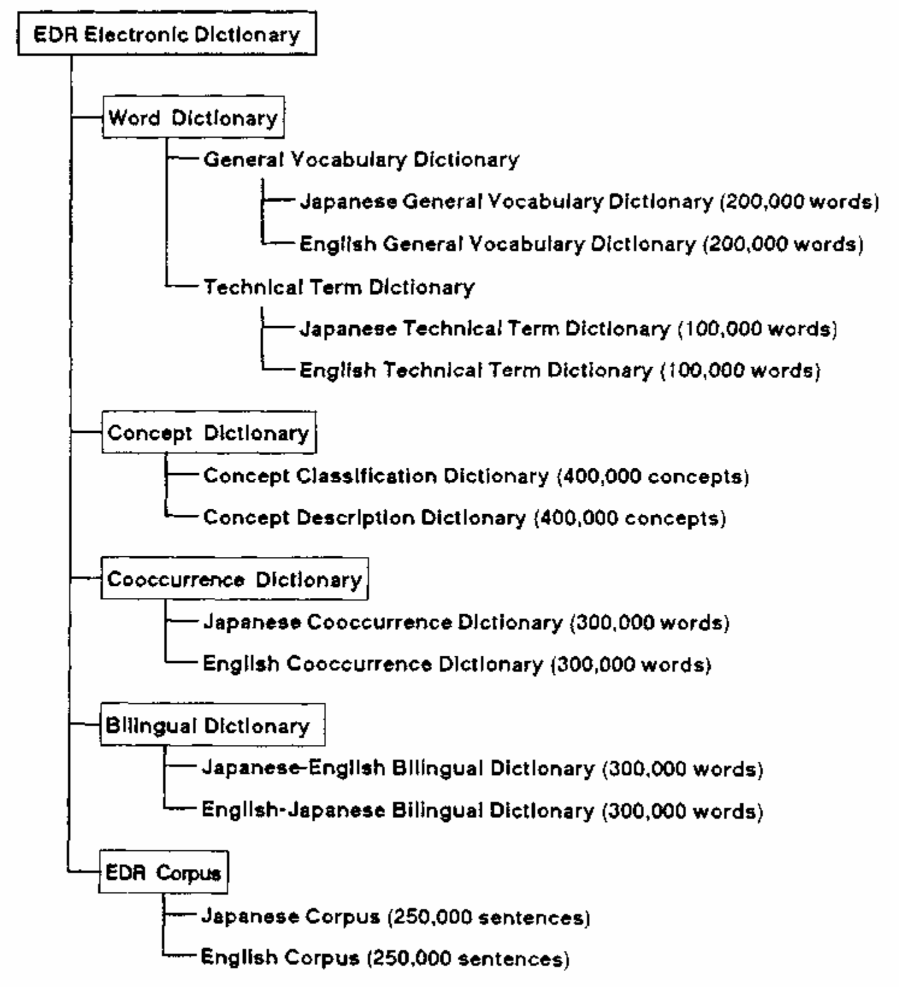
EDR Electronic Dictionary
Word Dictionary
General Vocabulary Dictionary
Japanese General Vocabulary Dictionary (200,000 words)
English General Vocabulary Dictionary (200,000 words)
Technical Term Dictionary
Japanese Technical Term Dictionary (100,000 words)
English Technical Term Dictionary (100,000 words)
Concept Dictionary
Concept Classification Dictionary (400,000 concepts)
Concept Description Dictionary (400,000 concepts)
Cooccurrence Dictionary
Japanese Cooccurrence Dictionary (300,000 words)
English Cooccurrence Dictionary (300,000 words)
Bilingual Dictionary
Japanese-English Bilingual Dictionary (300,000 words)
English-Japanese Bilingual Dictionary (300,000 words)
EDR Corpus
Japanese Corpus (250,000 sentences)
English Corpus (250,000 sentences)

Figure 2. Structure of the EDR Electronic Dictionary

attached to the concept identifier and used for users to distinguish one concept from others. The Concept Dictionary contains information on the 400,000 concepts defined in the Word Dictionary, and is divided into the Concept Classification and the Concept Description dictionaries depending on the type of information. The Concept Dictionary is necessary for the concepts listed in the Word Dictionary to be understood by computers. People use the knowledge they have already gained to understand sentences. In the same way, computers require knowledge base in order to understand written language. The Concept Dictionary provides this knowledge for computers. The concept, which is independent of languages, can be used in a universal intermediate language (interlingua) for machine translation, and the Concept Dictionary can be used in multilingual environments other than Japanese or English.

The Cooccurrence Dictionary is created for each language, and the Japanese and English Cooccurrence Dictionaries each contain 300,000 words. The Cooccurrence Dictionary contains information on wording which can be used when computers express a specific idea by choosing appropriate words in a natural language. This information is described by cooccurrence relations.

The Cooccurrence Dictionary is a collection of these cooccurrence relations.

The Bilingual Dictionary is divided by the direction of translation into Japanese-English and English-Japanese dictionaries, each of which lists 300,000 words including the general words and the technical terms. The Bilingual Dictionary provides information on the correspondences between Japanese and English headwords.

The EDR Corpus is comprised of the Japanese Corpus and the English Corpus, each containing 250,000 sentences. A large number of examples of use are provided to give linguistic data. The EDR Corpus is not a simple text base but contains syntactic and semantic structures of sentences. The EDR Corpus was created for use in developing the EDR Dictionary, but it can also be used in various types of research into natural language processing.

The development of the EDR Dictionary has been carried out according to the schedule shown in Figure 1. An outline of the current status of the dictionary is described below.

Word Dictionary: The development of four dictionaries (Japanese and English General Vocabulary Dictionaries of 200,000 words each, and Japanese and English Technical Term Dictionaries of 100,000 words each) has been concluded.

Concept Dictionary: The development of the Concept Classification for 400,000 concepts has been concluded. The Concept Description passed its first phase of development, and is currently in the improvement and expansion phase.

Cooccurrence Dictionary: This dictionary has been developed in connection with, or as a by-product of, the Concept Description, and therefore is in the same phase as the Concept Description development.

Bilingual Dictionary: This dictionary was developed on the theme of the Word Dictionary with the same schedule, that is, the development of the Bilingual Dictionary of 300,000 words is concluded.

EDR Corpus: The EDR Corpus has been developed on the theme of the Concept Description. The Japanese Corpus and the English Corpus of 250,000 sentences each have already been compiled. Currently, reviews and improvements of the syntactic and semantic analysis results in the Corpus are proceeding.

The dictionaries have been subjected to evaluations and improvements in the theme of the Evaluation System. In parallel with these internal evaluations, external evaluations by selected research groups are also introduced on a limited basis as a complement to the internal evaluations.

4. Towards International Cooperation

As mentioned above, in the ED Project, Japanese and English were chosen as the target languages. Although emphasis is naturally laid on Japanese, it is essential for the EDR Electronic Dictionary to include at least one language other than Japanese to prove the universality of the specifications of the Dictionary over various languages. English was adopted as a foreign target language because of its current status as a common international language. EDR has entrusted some of its tasks relating to English to the University of Manchester Institute of Science and Technology (UMIST) and the

Computing Research Laboratory (CRL) of the New Mexico State University (NMSU). Such cooperative works with research groups of native English speakers are indispensable to secure the quality of some parts of the EDR Electronic Dictionary relating to English. Considering the significant role that the English language plays in the world, it may be necessary for EDR to further broaden its spectrum of cooperative relations with English speaking people.

The ultimate goal of electronic dictionary projects should be to develop electronic dictionaries that can be applied to as many languages as possible throughout the whole world. EDR is currently making efforts to try to establish cooperative relations with research groups of native speakers other than English. For example, in Asia, EDR is cooperating with groups in China, Thailand, Malaysia and Indonesia through the CICC project. These groups are now developing their own electronic dictionaries based on the same specifications as those of EDR. EDR hopes that similar relationships will be established with other groups using different languages.

In cooperation with foreign groups, it is important to follow a procedure so that the cooperation results in equal benefits and the largest possible merits for both parties. The steps that are being taken by EDR are as follows:

Step 1: Exchange of information on technologies

The specifications of the EDR Electronic Dictionary and various other information which is obtained in the course of its research and development are to be available to the public as much as possible. The EDR Dictionary Interface, which is an entire set of concepts and headwords, has been made available to those who wish to know more about the content of the EDR Dictionary.

Step 2: Mutual cooperation on dictionary development

Further cooperation is possible with organizations which are planning to develop electronic dictionaries on specifications that are the same as or similar to those of EDR. EDR is ready for mutual presentation of detailed dictionary specifications and software for supporting systems of electronic dictionary development.

Step 3: Mutual exchange of electronic dictionaries

This is the final step of the cooperation. It may be possible to mutually exchange electronic dictionaries if the matter is fully discussed and agreed upon by both parties. Going through the process of these exchanges of dictionaries, a world where all groups can share the electronic dictionaries will be slowly but steadily realized. In this phase, the most important principle is that all languages should be treated on an equal basis.

5. Future Plans of the Project

The research and development of the EDR Electronic Dictionary will be continued up to the end of the project, that is, the end of the fiscal year of 1994, according to the schedule shown in Figure 1. Specifically, improvement and expansion of the Concept Description in the Concept Dictionary will be executed, and the Concept Description for 400,000 concepts will be completed. On the theme of the Evaluation System, full-function application systems for dictionary evaluation will be developed and dictionary evaluation using the evaluation systems will be undertaken. In the course of the dictionary evaluation, each dictionary of the EDR Electronic Dictionary will be evaluated upon various natural language interface systems, and evaluation results will be fed back to the dictionary for improvement.

All the final results of the ED Project will be distributed on a commercial basis after the project terminates. However, it is true that some parts of the present EDR Dictionary have come to a state in which they can be used for experimental purposes, and a number of serious requests for using them as early as possible for their research purposes have been made to EDR. Therefore the intermediate results should also be provided as a part of field tests for those who want to use them. EDR is now preparing a scheme to offer the intermediate results on a joint research basis, on the expectation that this scheme will be realized before long.

EDR's fundamental policy of treating the results is summarized below:

(1) The same conditions regarding the usage of the EDR Electronic Dictionary will be applied to all users no matter whether they are domestic or overseas users.

(2) The prices will be set somewhat lower than those of machine-readable dictionaries that are currently on sale.

(3) Special measures will be arranged for those users for academic purposes, such as universities and public research institutions.

In general, electronic dictionaries should continuously be improved and expanded after completion of the initial development. Without efforts for improving and expanding dictionaries, users could not rely upon the dictionaries. The efforts, however, involve a problem of bearing the expense for maintaining the dictionaries.

EDR recognizes that it is also indispensable for the EDR Dictionary to be maintained continuously after the termination of the project, and that the maintenance cost should be covered with an income
from the distribution of the Dictionary. Furthermore, to continue the dictionary maintenance over a long period, it is necessary to establish a firm system or organization. Discussion of these issues has started. We expect before long to come to a viable conclusion.

Note:

1. Fujitsu, Ltd., NEC Corporation, Hitachi, Ltd., Sharp Corporation, Toshiba Corporation, Oki Electric Industry Co., Ltd., Mitsubishi Electric Corporation and Matsushita Electric Industrial Co., Ltd.