An Overview of the EDR Electronic Dictionary and the Current Status of Its Utilization

Hideo Miyoshi, Kenji Sugiyama, Masahiro Kobayashi, and Takano Ogino
Japan Electronic Dictionary Research Institute, LTD. (EDR)
Daini-Abe Bldg., 78-1, Kanda-Sakumagashi, Chiyoda-ku, Tokyo 101, Japan
{miyoshi, kenji, kobayasi, ogino}@edr.co.jp

Abstract

In this paper we present the specification and the structure of EDR Electronic Dictionary which was developed in a nine-year project. The first version of EDR dictionary (V1.0) and its revised version (V1.5) are already released and are now utilized at many sites for both academic and commercial purposes. We also describe the current status how the EDR dictionary is utilized. Finally we will give the outline of the new R&D project which EDR will launch in fiscal 1996.

1 Introduction

The EDR Electronic Dictionary[1,2,3] is the result of a nine-year project (from fiscal 1986 to fiscal 1994), funded by the Japan Key Technology Center and eight computer manufacturers* aimed at establishing an infrastructure for advanced processing of natural language by computers and knowledge information processing.

The features of the EDR Electronic Dictionary can be summarized as follows:

(1) A large scale that covers all the vocabulary used in ordinary writing
(2) Aimed at general purpose applications without bias towards a particular application system or algorithm
(3) Provided with the knowledge base required for true semantic analysis
(4) A high degree of objectivity based on large volumes of text
(5) Fundamental content that is highly generalized across different languages and fields

The EDR Electronic Dictionary, which is composed of eleven sub-dictionaries, catalogues the lexical knowledge of Japanese and English (the Word Dictionary, the Bilingual Dictionary, and the Co-occurrence Dictionary), and has unified thesaurus-like concept classifications (the Concept Dictionary) with corpus databases (the EDR Corpus). The Concept Classification Dictionary, a sub-dictionary of the Concept Dictionary, describes the similarity relation among concepts listed in the Word Dictionary. The EDR Corpus is the source for the information described in each of the sub-dictionaries. The basic approach taken during the development of the dictionary was to avoid a particular linguistic theory and to allow for adoptability to various applications.

The first version of EDR dictionary (V1.0) and its revised version (V1.5) are already released and are now utilized at many sites for both academic and commercial purposes. This paper outlines the specification of EDR Electronic Dictionary and describes the current status of its utilization.

2 The Structure of the EDR Electronic Dictionary

The EDR Electronic Dictionary is composed of five types of dictionaries (Word, Bilingual, Concept, Co-occurrence, and Technical Terminology), as well as the EDR Corpus.

EDR Electronic Dictionary
- Word Dictionary
  - Japanese Word Dictionary
  - English Word Dictionary
- Bilingual Dictionary
  - Jpn.-Eng. Bilingual Dictionary
  - Eng.-Jpn. Bilingual Dictionary
- Concept Dictionary
  - Headconcept Dictionary
  - Concept Classification Dictionary
  - Concept Description Dictionary
- Co-occurrence Dictionary
  - Japanese Co-occurrence Dictionary

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

    - English Co-occurrence Dictionary
- Technical Terminology Dictionary
  - Jpn. Technical Terminology Dictionary (Information Processing)
  - Eng. Technical Terminology Dictionary (Information Processing)
  - Others (Concept Classification, Bilingual Dictionary, Co-occurrence Data, etc.)
- EDR Corpus
  - Japanese Corpus
  - English Corpus

The Japanese Word Dictionary contains 250,000 words, and the English Word Dictionary contains 190,000 words.

The Bilingual Dictionary lists the correspondences between headwords in the different languages. The Japanese-English Bilingual Dictionary contains 230,000 words, and the English-Japanese Bilingual Dictionary contains 190,000 words.

The Concept Dictionary contains information on the 400,000 concepts listed in the Word Dictionary and is divided according to information type into the Headconcept Dictionary, the Concept Classification Dictionary, and the Concept Description Dictionary. The Headconcept Dictionary describes information on the concepts themselves. The Concept Classification Dictionary describes the super-sub relations among the 400,000 concepts. The Concept Description Dictionary describes the semantic (binary) relations, such as 'agent,' 'implement,' and 'place,' between concepts that co-occur in a sentence.

The Co-occurrence Dictionary describes collocational information in the form of binary relations. The Japanese Co-occurrence Dictionary contains 900,000 phrases, and the English Co-occurrence Dictionary contains 460,000 phrases.

The Technical Terminology Dictionary covers the field of information processing, and is split into four types of dictionaries of Word, Bilingual, Concept (Classification), and Co-occurrence.

The linguistic data which the EDR Corpus contains has been obtained by collecting a large number of example sentences and analyzing them on morphological, syntactic, and semantic levels. The Japanese Corpus contains 220,000 sentences, and the English Corpus contains 160,000 sentences.

3 Role of Each Dictionary

This chapter describes the roles of the major subdictionaryies of the EDR Electronic Dictionary and shows some examples.

3.1 Word Dictionary

The role of the Word Dictionary is to provide part of the information on the morphological, syntactic, and semantic revels that is required for natulal language processing. Morphological information relates to headword (morpheme) and information on the connectivity of morphemes. This is used in morphological analysis to find the morphemes, and also used in morphological generation to produce output sentences.

Information on the syntactic level includes parts of speech as well as surface case information and other grammatical attributes. This information is used in syntactic analysis and generation, and provides the basis for the formulation of parsing rules and production rules.

Semantic information includes concept identifiers. Headconcept and concept explications are provided as accampanying information. The concept identifier is a numerical expression and the basic constituent of the Concept Dictionary. The headconcept is a representative word that is the most appropriate in expressing the concept identified by the concept identifier. The concept explication is an explanation written in natural language for the purpose of assisting humans in differentiating one concept from another. Every Word Dictionary record has a concept identifier to link the Word Dictionary and the Concept Dictionary.

The following is an example of English Word Dictionary record:

```
Headword: dog
Connectivity: ELN1, ECN1
Part of Speech: EN1 (common noun)
Grammatical Attributes: ECN1;ENSG, ENC;ENNE
Concept_ID: 3dbc67
Headconcept: dog
Concept Explication: an animal called dog
```

3.2 Bilingual Dictionary

The Bilingual Dictionary is designed to give appropriate correspondence words to the headwords contained in the Word Dictionary, in machine processings. The headword information of the Bilingual Dictionary is a subset of the Word Dictionary, that is, headword notations, parts of speech, concept identifiers, headconcepts, and concept explications. The concept identifiers and concept explications are used to indentify the meaning of the polysemous headwords. Some of the correspondence words include additional information which describes the constraints where the correspondence words are used.

The following is an example of English Japanese Biligual Dictionary record:

```
Headword: dog
Part of Speech: EN1 (common noun)
Concept_ID: 3dbc67
Headconcept: dog
Concept Explication: an animal called dog
Correspondence Word: 犬
```

3.3 Concept Dictionary

The role of the Concept Dictionary is to provide the data required for computer processing of the semantic contents or the concepts, expressed in natural language sentences, such as:

(1) Generating appropriate semantic representations for sentences

(2) Determining the similarity (equivalence) of semantic contents

(3) Converting a semantic content into a similar (equivalent) content

For this reason, the Concept Dictionary contains three types of subdictionaries: Headconcept Dictionary, Concept Classification Dictionary, and Concept Description Dictionary. In the Concept Dictionary, each concept is uniquely identified by a concept identifier which is a hexadecimal number. The Headconcept Dictionary contains the concept identifier and the headconcept, and the concept explication. The headconcept is a word whose meaning is close to the content meaning of the concept. The concept explication is an explanation which expresses the meaning of the concept. The Concept Classification Dictionary contains the set of pairs of concepts that have super-sub (is_a) relation. For example, the super-concepts of 'school' are 'organization,' 'building,' and 'function.' The sub-concepts of 'school' are 'elementary school,' 'university,' and so forth. The Concept Description Dictionary contains the set of pairs of concepts that have certain semantic relations other than super-sub relations. The following eight semantic relations are used:

object agent goal implement
a-object place scene cause

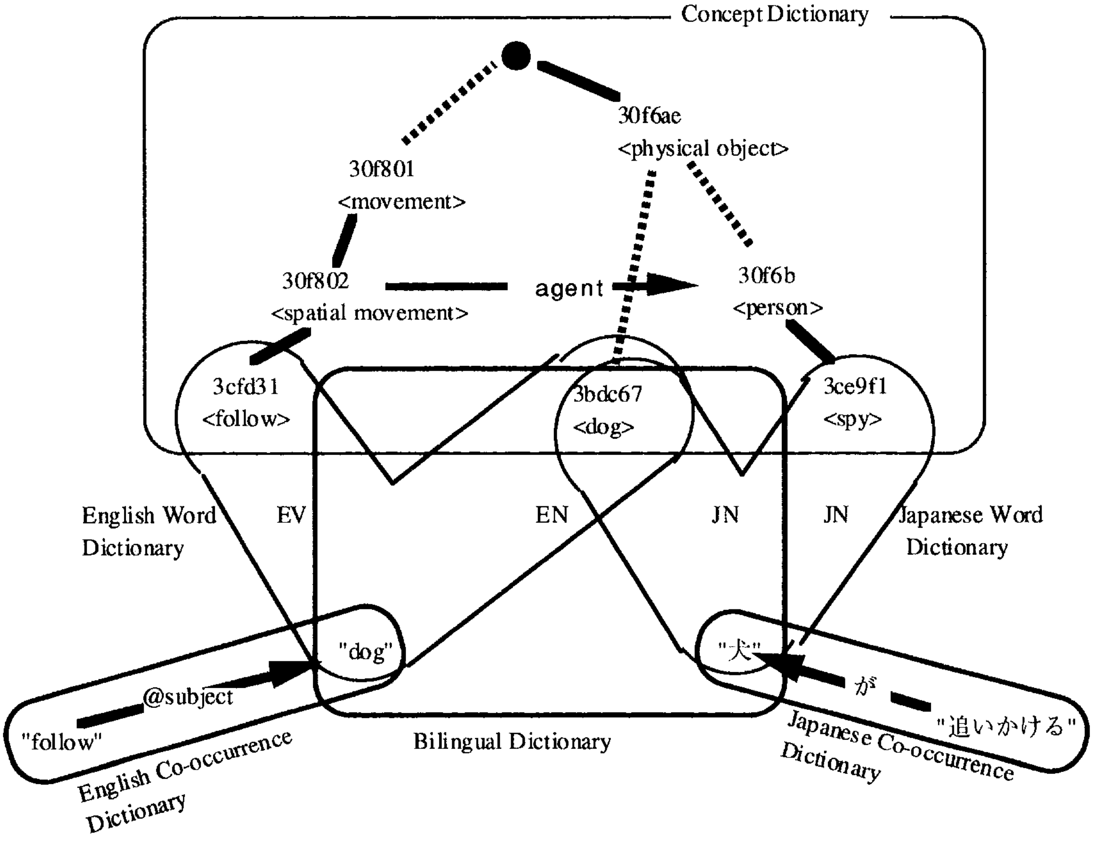

Figure 1. Relationships between sections of EDR Electronic Dictionary

Table 1: Number of User Sites of the EDR Electronic Dictionary

	university			goverment institution	private company	total
	national and public	private	overseas			
No. of User Sites	48	18	1	3	23	93

3.4 Co-occurrence Dictionary

The Co-occurrence Dictionary includes the type of word conbinations used to construct a sentence, that is, collocational information. This type of information is used to select the appropriate correspondence words in machine translation.

3.5 EDR Corpus

The EDR Corpus is composed of the record number, sentence information, constituent information, morphological information, syntactic information, and semantic information. The basic role of the EDR Corpus is first to identify the sentence constituents of sentences, and then to indicate how the constituents combine to form the morphological, syntactic and semantic structure of the sentence using a large number of actual examples. The data in the Concept Description Dictionary and the Co-occurrence Dictionary is extracted from the EDR Corpus.

These subdictionaries are not indendent, but are organically connected (Figure 1).

4 The Current Status of Utilization

As we mentioned in chapter 1, we have already released the first CD-ROM version of EDR Electronic dictionary (V1.0) in April 1995 after the nine year R&D project. They are now being utilized at many sites for both academic and commercial purposes (Table 1). In fiscal 1995, furthermore refinement and improvement were done and the revised version (V1.5) is available since April 1996. One of the users, Fujitsu, released a commercial product using the EDR Electronic Dictionary in 1995. The product is called "Denjikai for Windows V2.0," which retrieves the word information from various dictionaries including EDR Electronic Dictionary.

5 Conclusion and Future Work

A number of dictionaries are currentry being developed under the name of electronic dictionaries (machine-readable dictionaries). These dictionaries consist of information from published dictionaries that has been stored on a recording medium, and which can then be referred to and used by mechanical means. However, these electronic dictionaries are referred to and used by people, unlike true electronic dictionaries (machine-tractable dictionaries), which in the strict sense are intended for use in machine processing. True electronic dictionaries are not simply machine-readable editions of dictionries for use by people. They must include all the information necessary for a computer to understand a natural language. We think that the EDR Electronic Dictionary satisfies those conditions and hope that it will be widely used for various natural language processing applications.

Finally we would like to make a short remark on the new project which EDR will launch in fiscal 1996. The new project will be funded by Information Technology Promotion Agency (IPA) of Japan and will be carried out in conjunction with Tokyo Institute of Technology and Tokyo University. The objective of the project will be the creation of a software that will allow the linguistic knowledge base to automatically expand by feeding the output of analyzed text into the knowledge base itself. We hope this will help refine and extend the EDR Electronic Dictionary.

References

[1] EDR, Proceedings of the International Workshop on Electronic Dictionaries, EDR TR-031, 1991.

[2] EDR, EDR Electronic Dictionary Version 1 Technical Guide, EDR TR2-003, 1995.

[3] EDR, Summary for the EDR Electronic Dictionary Version 1 Technical Guide, EDR TR2-005, 1995.