A Mono-lingual Corpus-Based Machine Translation of the Interlingua Method

Eiji KOMATSU, CUI Jin and Hiroshi YASUHARA

Japan Electronic Dictionary Research Institute Ltd.
C/O Systems laboratory, OKI Electric Industry Co., Ltd.
11-22, Shibaura 4-chome, Minato-ku, Tokyo 108 Japan
e-mail: komatsu{sai, yasuhara }%edr6r.edr.co.jp@uunet.uu.net

Abstract

This paper describes a prototype of an example-based machine translation system. In this system, key language resources are EDR corpus and concept classification dictionary. The corpus consists of a pair of sentences, their morphological representations, their syntactic representations, and their semantic representations. The semantic representations are described by an interlingua. Therefore the corpus can be viewed either as a mono-lingual corpus or as a parallel corpus between a natural language and an interlingua. The system analyses source sentences and generates target sentences by example databases. Similarity calculations play essential roles in analysis and generation phases. These calculations uses the concept classification dictionary. The translation system is realized by directly combining a source language analysis and a target language generation without a transfer phase. The system has been implemented and the state of the current prototype showed evaluation data which suggested the corpus-based MT approach would be good prospects.

1. Introduction

Many machine translation systems have been developed in the last decade, and some of them have become commercialized. However, the qualities of these systems is not sufficient enough for practical usage. Almost all of these systems can be grouped into rule-based systems, and they seem to have reached the limitations of the method. On the other hand, there is research trend that is returning to the study of the human process of translation and is trying to introduce the human mechanism of translations into the machine translation systems [8]. "Corpus-based" or "example-based" machine translations are included in this trend and many methods of this line have been proposed [1],[6],[7],[8],[9],[10],[11]. However, an Example-Based MT system needs a large quantity of good examples, and a language knowledge base to help retrieval and use similar examples in the example database.

EDR has been developing large-scale dictionaries: Word dictionaries (Japanese and English), Bilingual dictionaries (Japanese-to-English and English-to-Japanese), a Concept dictionary,

Co-occurrence dictionaries (Japanese and English), and EDR corpora (Japanese and English mono-lingual corpora that consist of raw text, morphological alignments, syntactical trees and interlingua). The corpus data will consist of as many as 250,000 Japanese sentences and 250,000 English sentences upon completion.

Since Example-Based MT systems need a large number of examples and EDR is making large-scale corpus data, this paper will deal with a method making use of the EDR corpora as an elements of Example-Based MT systems. Because the EDR corpora are mono-lingual and include interlingua [9], it is convenient to use them in semantic analysis or semantic generation of natural language. The EDR dictionaries are used as the language knowledge base in our Example-Based MT system.

This paper will emphasize that the example-based technique is effective for MT that uses interlingua and large-scale corpora. The EDR Electronic are not discussed in detail here. The system is restrained to the translation between Japanese and English syntactic trees and interlingua. The term "semantic" is used to mean the translation between syntactic trees and interlingua.

Chapter 2 explains the EDR Electronic Dictionaries we used and chapter 3 is an explanation of the system overview. In chapters 4 and 5 the Japanese semantic analysis and the English semantic generation are discussed. Chapter 6 shows the results of the experiments and chapter 7 summarizes those results.

2. EDR Electronic Dictionaries

The EDR electronic dictionaries consist of word dictionaries, a concept dictionaries, co-occurrence dictionaries, and bilingual dictionaries. Figure 2.1 shows the structure of EDR Electronic Dictionaries. This system uses the Word Dictionaries, the Concept Dictionaries and the EDR Corpora.

2.1 Word Dictionaries and Concept Dictionary

The information in the Word Dictionaries are 1) PCS (Part Of Speech), 2) adjacent information, 3) grammatical information and 4) a concept represented by a word.

The EDR Concept Dictionary [2] represents a variety of knowledge related to the concepts carried by words (meanings of words) in a form understood by a computer. The number of concepts in the EDR concept dictionary is 400,000. Concepts are represented by numbers called concept identification numbers (referred as "ID" below). The Concept Dictionary is divided into concept descriptions and concept classifications by relation types. Only the concept classification is used in this paper. Here, information not used are omitted.

Concept classifications provide a hierarchy of concepts created to reduce the volume of knowledge described by enabling inheritance of knowledge from super-concepts to sub-concepts. Concepts with common attributes are grouped together and are classified based on a super-concept assigned to each group. One concept belongs to various groups because each group is formed focussing on a separate attribute. That is, a concept can be defined as inheriting multiple super-concepts, and the sub-concept can be regarded as a set of attributes by which its super-concepts are defined.

Figure 2.2 shows part of an EDR concept classification.

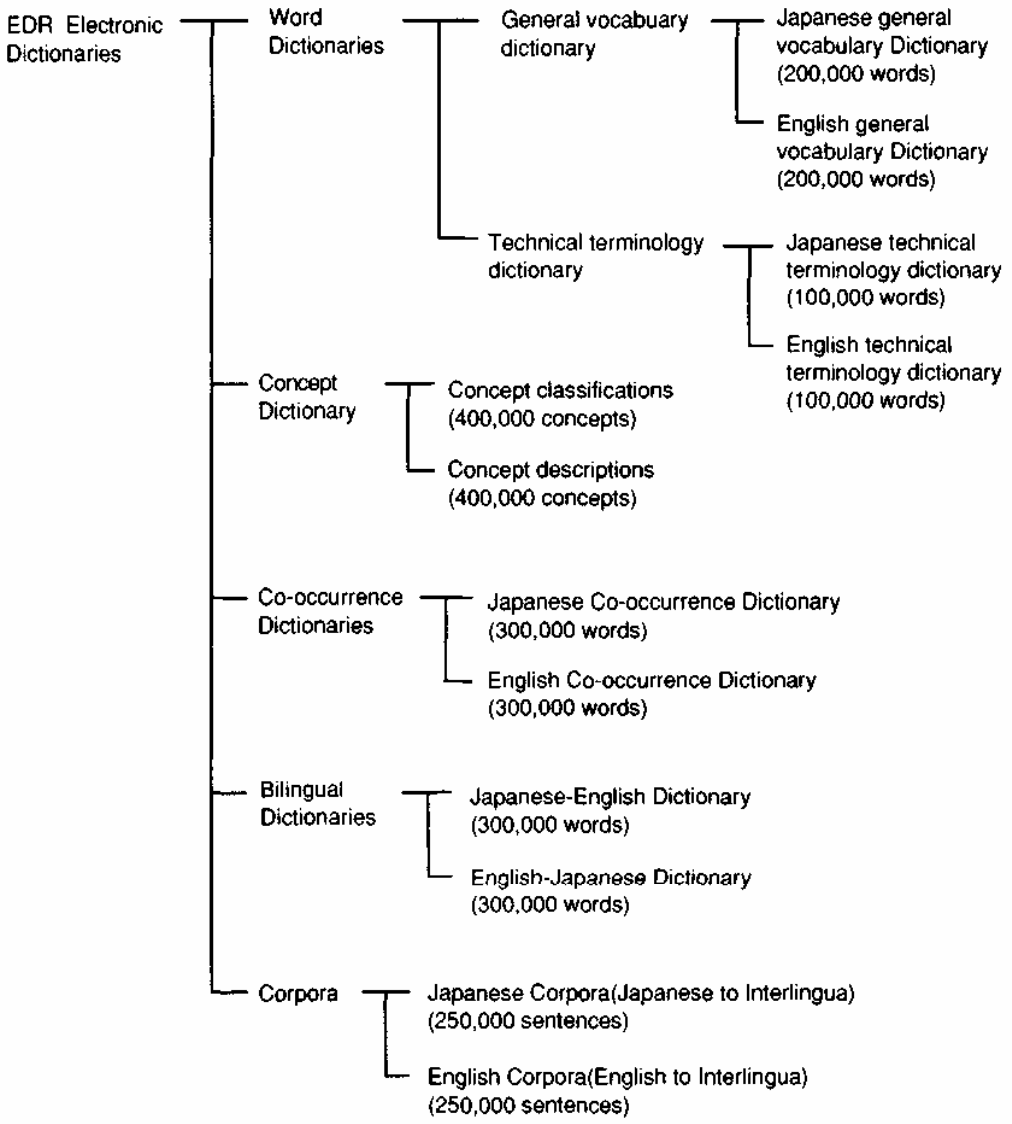


Figure 2.1: Structure of EDR Electronic Dictionaries

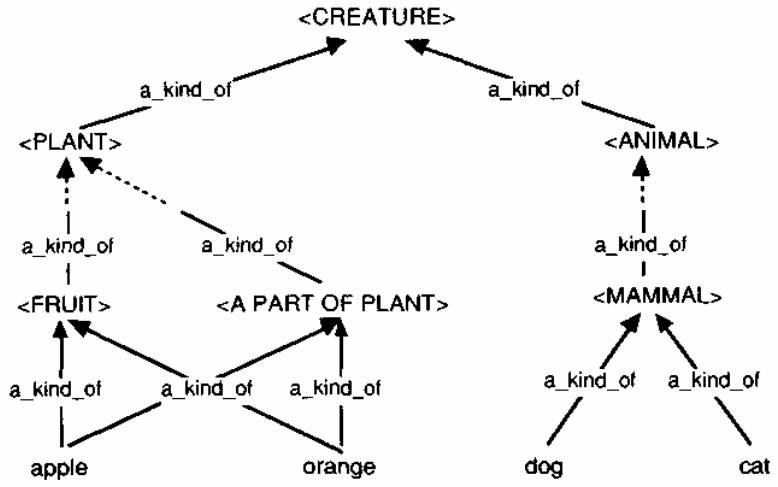


Figure 2.2: EDR concept classifications

2.2 EDR Corpora

For the dictionary development in EDR, a large-scale Japanese corpus and an English corpus are being made. EDR corpora [3] are mono-lingual, and the sentences in the corpora database are selected from the EDR text database, which consists of 20,000,000 Japanese sentences and 20,000,000 English sentences. The format of the corpora are basically the same for both Japanese and English. An EDR corpus consists of 1) raw text, 2) morphological alignment, 3) syntactical tree and 4) interlingua.

Figure 2.3 shows EDR corpus data that is the Japanese sentence "彼は面白い本を読む(*He reads the interesting book.)* . Numbers preceding IDs mean that the concept and the word corresponds.

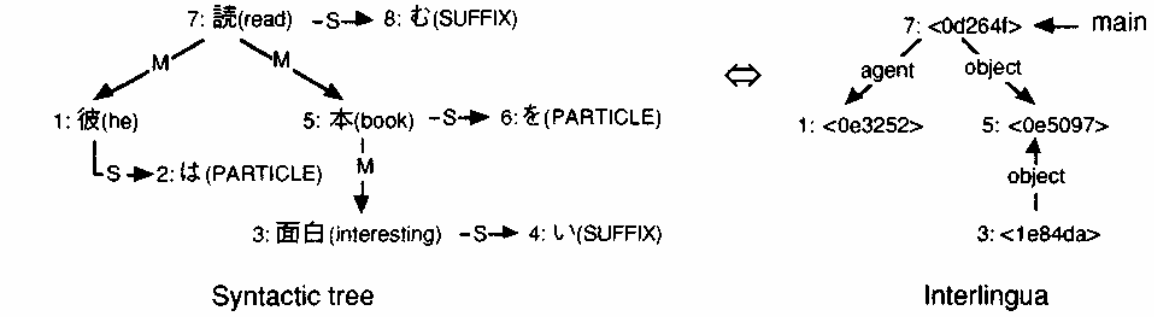


Figure 2.3: EDR Corpora (Japanese)

In figure 2.3, the numbers in <> express the ID of a concept.

<0e3252> : a demonstrative personal pronoun
<1e84da> : interesting
<0e5097> : publications
<0d264f> : to get the stated information from print or writing

In syntactic trees, there are two kinds of relations. "M" relation means the relation between content words (e.g. noun, verb, adjective, etc). "S" relation means the relation between a content word and a function word(e.g. particle, suffix, etc). In interlingua, relations between concepts are defined by EDR [2]. Both interlingua and syntactic trees are dependency representations. Nodes of interlingua are concepts, and nodes of syntactic trees are words.

In this paper, there are two assumptions for corpus data: 1) correspondences between words and concepts should be one-to-one and 2) each concept relation should appear only once for one concept in an interlingua.

Figure 2.4 shows EDR corpus data.

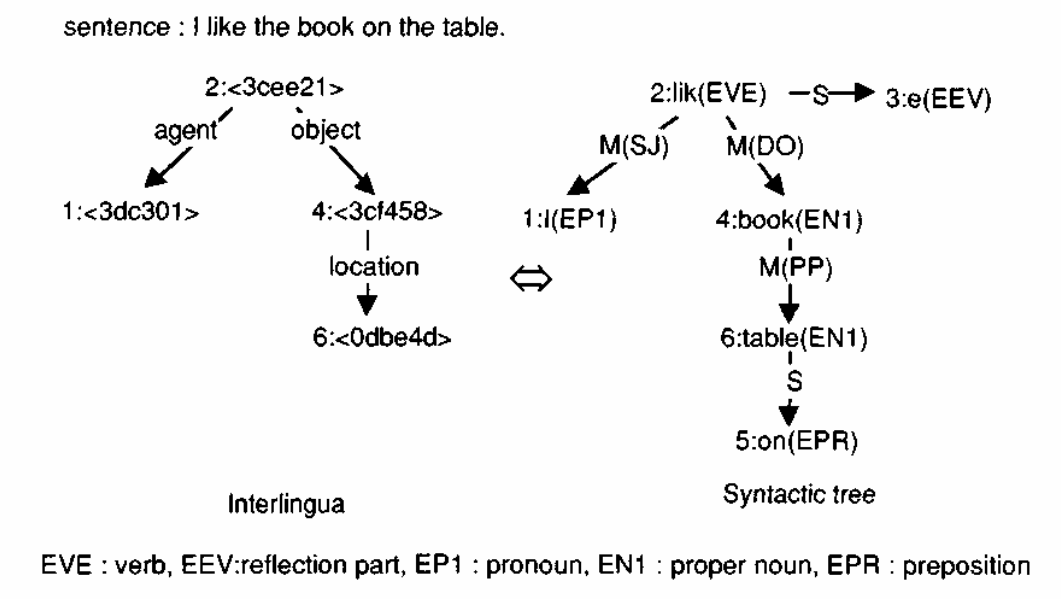


Figure 2.4: EDR Corpora (English)

In Figure 2.4, the numbers in <> express the ID of a concept.

<3cee21> : to like something
<3dc301> : c#I
<3cf458> : a set of written, printed or blank sheets bound together into a volume
<0dbe4d> : a table upon which food is served

In this corpus, subcategories of the "M" relations for the generator are made as follows:

SJ: subject	obligatory
DO: direct object	obligatory
IO: indirect object	obligatory
SC: subjective complement	obligatory
OC: objective complement	obligatory
PP: prepositional phrase	optional
ADV: adverbial word	optional
ADJ: adjectival word	optional
HD: head node	

3. System Overview

The system is 1) a Corpus-based machine translation system; 2) the EDR dictionaries (including word dictionaries, concepts dictionaries and large-scale mono-lingual corpora) used as knowledge resources; 3) a use of the pivot method, the target for analysis and the source for generation are Interlingua based on EDR concept dictionary.

Because some of the technology of morphological analysis/generation and syntactic analysis /generation [12] have been proven to be effective and some of them have become commercialized, this system is restricted to the translation between syntactic representations of Japanese and English through semantic representations (Interlingua). It is well known that the processing of meaning is one of the essential and difficult tasks in MT systems or other natural language processing systems.

Figure 3.1 shows the configuration of our machine translation system. Translating Japanese syntactic trees to an English syntactic tree is carried out by the Japanese semantic analyzer and the English semantic generator. The Japanese semantic analyzer takes a Japanese syntactic tree and transfers it to an interlingua, the English generator transfers the interlingua to an English syntactic tree.

The knowledge resources of our system is the EDR dictionaries and example databases created from EDR corpora. The grammatical rules are not used.

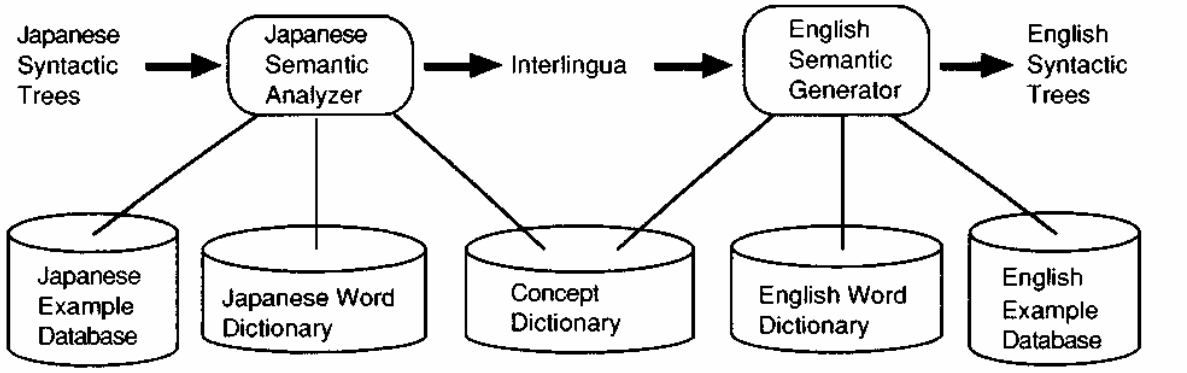


Figure 3.1: The configuration of the Corpus-based machine translation system

4 Japanese Semantic Analysis

The semantic analysis in this paper means determining the word sense of every word and obtaining an appropriate semantic structure, when a syntactic structure of a sentence is inputted. Since the EBMT systems need a large quantity of examples and EDR is making large-scale corpus data, in this chapter, deals with a method making use of the EDR corpora as examples of EBMT systems. Because the EDR corpora are mono-lingual and include interlingua [14], it is more convenient to use them in semantic analysis or semantic generation of natural language. This system differs from preceding Example-Based systems in two ways: 1) In Japanese analysis, the similarity between words (distance between words) is calculated based on their concept (word sense) using the EDR Electronic Dictionaries; and 2) the examples in our system are mono-lingual.

4.1. Examples Created from EDR Corpora

It seems unwise to use the corpora data directly because there are many long and complex sentences in them. Since it is intended to make use of the expansiveness of the corpora rather than the detail of it, the corpora date was divided into several elementary units and used as examples. In other words, a complex sentence was divided into simple sentences and phrases.

An example in this Example Database consists of a pair of a syntactic tree and an interlingua,

together with correspondences between words in the syntactic tree and concepts in the interlingua. Syntactic trees and interlinguae are both dependency representations. The syntactic tree consists of words and surface relations. The interlingua consists of concepts and concept relations.

Figure 4.1 shows two examples created from the corpora data shown in Figure 2.3.

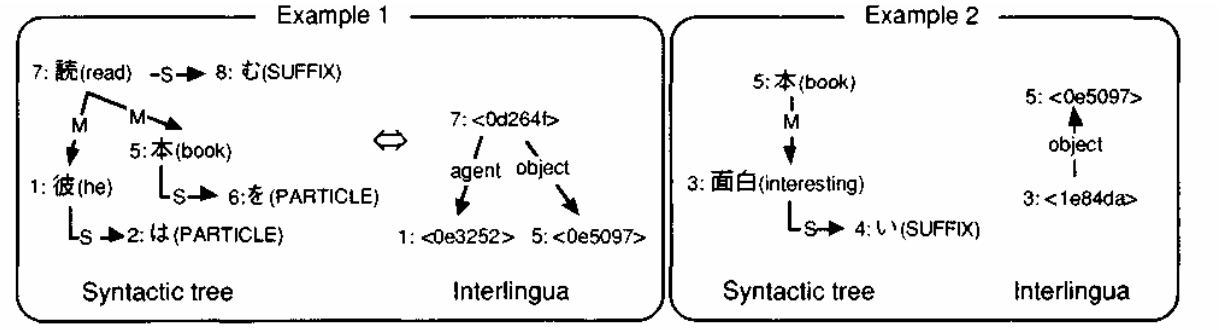


Figure 4.1: Examples created from one corpus data

All of the examples have a head word and some surface relations related to the head word. There are two kinds of examples. One is a simple sentence whose head is a predicate such as a verb or adjective; the other is a phrase whose head is nounal word such as a noun or pronoun.

Japanese verbs can be used as sample data to illustrate the coverage of examples created from the EDR corpora. In the EDR Japanese Dictionary, the number of Japanese verbs is as many as 43,891 words. The total number of concepts used by Japanese verbs is 29,676 concepts. One of the aims is to make more than one example for every Japanese verb concept in order to do semantic analysis of Japanese simple sentences. Therefore, the appearance of 29,676 concepts of Japanese verbs on 12,000 EDR corpora was investigated.

corpora	verbs that appear in corpora(percentage)
12,000	2,926 (9.8%)

Figure 4.2: the coverage of EDR corpora

Upon completion of the project, it is hoped the EDR corpora (250,000 sentences) will cover nearly all of the Japanese verb concepts. For the verb concepts that do not appear in the EDR corpora, it is planned to collect them from an online text, and then add them to the Example Database and the EDR corpus set.

4.2 Similarity Calculation

There are three kinds of similarities that need to be calculated in this system: 1) The similarity between two words; 2) The similarity between two concepts; 3) The similarity between two simple sentences or phrases. All of these are based on the EDR Word Dictionary [4, 5] and Concept Dictionary.

4.2.1 The Similarity Between Words

In the system the similarities between two words is calculated based on their concepts using

the EDR Word Dictionary and Concept Dictionary [1].

Suppose W1 and W2 are two words. The similarity between W1 and W2 is calculated based on two kinds of similar relations between them. One is a similar relation, we call relation a. between the concept set of W1 and concept set of W2. The other is a similar relation, we call it relation β, between the super-concept set of W1 and the super-concept set of W2 .

Suppose the C1 is the number of concept sets for word W1, and the C2 is the number of concept sets for word W2 . | X| expresses the number in the set X, then the similarity α is:

$$\alpha = | C_1 \cap C_2 | \qquad (4.1)$$

In order to calculate the similarity β, the super-concepts are retrieved from each concept in C1 and C2 . Supposing CS_κ is the number of common concepts between the super-concept set of W1 and the super-concept set of W2 in the super-concept level K. $N_{\kappa 1}$ is the number of the super-concept of W1 in the super-concept level K, and $N_{\kappa 2}$ is the number of the super-concept of W2 in the super-concept level K. Then the similarity between W1 and W2 in the super-concept level K, called β_κ , is:

$$\beta_\kappa = (1 + K_{\beta 1} * CS_\kappa)(1 + K_{\beta 2} * (\frac{CS_\kappa}{N_{\kappa 1}} + \frac{CS_\kappa}{N_{\kappa 2}})) - 1 \qquad (4.2)$$

$K_{\beta 1}$ is used to adjust the weight of CS_κ; $K_{\beta 2}$ is used to adjust the weight of $CS_\kappa/N_{\kappa 1}+CS_\kappa/N_{\kappa 2}$.

The similarity β is calculated based on each β_κ using the following formula. The number N can be specified case by case, as suits the user's needs.

$$\beta = K_1 * \beta_1 + K_2 * \beta_2 + ... + K_N * \beta_N \qquad (4.3)$$

K_i { i = 1, 2 ... N } are used to adjust the weight of each similarity β_κ .

The similarity between the two words W1 and W2 , called similarity σ, is calculated by using the following formula :

$$\sigma = similarity\_of\_word (W_1, W_2) = 1 - e^{-(K_\alpha * \alpha + K_\beta * \beta)} \qquad (4.4)$$

The similarity σ takes the value from "0" to "1". The nearer to "1", the more similarity there is between the two words W1 and W2. K_α can be used to adjust the weight of similarity α, K_β can be used to adjust the weight of similarity β.

The function *similarity_of_word (* W1, W2) can also be used to calculate similarity between two words that belong to two different languages because our similarities are calculated based on the concepts of words instead of words themselves. The concept dictionary is not dependent on a type of language.

The values of weight in formulas (4.2), (4.3) and (4.4) depend on the structure of the concept dictionary being used. They are determined and adjusted according to the structure of the concept dictionary. Figure 4.3 is a group of values of weight .

K_α	K_β	$K_{\beta 1}$	$K_{\beta 2}$	K_1	K_2
0.45	0.028	2.75	8.25	1	0.05

Figure 4.3 : values of weight (N=2)

Using the values of weight shown in Figure 4.4 and the EDR word dictionary and concept dictionary, some calculated similarities are shown in Figure 4.4 .

W1	W2	similarity
dog	dog	1
犬(*dog*)	cat	0.7423
犬(*dog*)	猫(*cat*)	0.8380
apple	orange	0.7646
dog	apple	0
country	son	0.0234
last	remain	0.7442
continue	remain	0.7771
black	white	0.9695

Figure 4.4 : Similarities between two words

4.2.2 The Similarity Between Concepts

The similarity between two concepts depends on the number of common super-concepts and the number of super-concepts. Formulas (4.2), (4.3) and (4.4) can also be used to calculate the similarity between concepts. But in this case, the CSk is the number of common concepts between the super-concept set of Concept1 and the super-concept set of Concept2 in super-concept level K. Nk1 is the number of super-concepts of Concept1 in super-concept level K, and Nk2 is the number of super-concepts of Concept2 in super-concept level K.

4.2.3 The Similarity Between Sentences or Phrases

It is clear that the similarity between two sentences mainly depends on their syntactic structure and the key words that appear in each sentence. In this system, the following two factors are used to determine the similarity between two simple sentences. One is the head, which is usually the main verb to a simple sentence or the modified side in a phrase; the other is the surface relations, that are pairs [F, W] (F is a function word and W is generally a noun), related to the main verb.

The similarity between two simple sentences, expressed as SIMsen, is calculated following the procedure shown below:

1) Calculating the SIM_{HEAD}, that is, the similarity between two head words, HEAD1 and HEAD2.

$$SIM_{HEAD} = similarity\_of\_word(HEAD1, HEAD2)$$

2) Searching the pairs with the same function words in two sentences. The similarity between the two pairs [Fi, Wi1] and [Fi, Wi2] is expressed as "SIMi", i (i = 1,2, ... n) is a progressive number that starts from "0" for a phrase or a simple sentence. The similarity of words Wi1 and Wi2 are calculated.

SIMi = *weight* (Fi) * *similarity_of_ word* (Wi1, Wi2)

weight (Fi) is used as the function to calculate the weight of the function word Fi. Among all Japanese function words, some play more important roles in determining the similarity of two sentence, such as "wo(を)", "ga(が)", "he(へ)", "ni(に)", "kara(から)", "made(まで)", They should be weighted more than the others.
Supposing the value of normal function words is "1", the weight of these function words in our system is shown in Figure 4.5.

function word	weight
wo(を)	2.75
ga(が)	2.25
he(へ)	3
ni(に)	1.2
kara(から)	1.85
made(まで)	1.5

Figure 4.5: The weights of some function words

3) The similarity between two sentences is calculated based on the SIMHEAD and SIMi calculated in step 1) and 2) using the following formula :

SIMSEN= SIMHEAD + Σ SIMi

The similarity between two phrases can be treated in the same way described here. But instead of the main verb, the head of a phrase is the modified side.

4.3 Semantic Analysis Based on Examples

In order to determine the meaning of every word and obtain an appropriate concept structure to a input SYN, the following procedure is used to search and refer to similar examples in the Examples Database:

1) Retrieving examples using the head word and its synonymy.
   Retrieving examples from the Example Database using the head word and the synonymy *as* a retrieving key.
2) Grouping the examples that have a similarity larger than the assigned value.
   Calculating the similarity between the input and examples retrieved, grouping the examples that have a similarity larger than the value assigned by the user, we express this group as G={ E1, E2, ...EN }, Ei { i = 1,2,… N }
3) Determining the concept of the head word.
   Finding the example Ek that has the largest similarity in group G, determining the concept of the head word in Ek as the concept of the head word of the input.
4) Surface relation analysis 1: finding the most similar surface relation.
   To a surface relation [FINj, WINj] in the input, searching G to find a surface relation [FEX, WEX], FINj =FEX and *similarity_of _word (* WINj, WEX), which has the largest similarity.
5) Surface relation analysis 2: determine the concept of the word in surface relation.

If *similarity_of_word* (WINj , WEX) is larger than the the value assigned by the user, calculating the similarity between the concept of WEX and each concept of WINj , determine the concept of WINj which has the greatest similarity to the concept of WINj.

6) Surface relation analysis 3: determine the concept relation of surface relation and head.
   The concept relation between the concept of the head word and the concept of WINj is determined by referring to [FEX, WEX] .
7) Back to 4)
   Repeat 4) 5) 6) until all surface relations in the input are analyzed.

Suppose there are two related examples in the Example Database. One is "鷹が飛ぶ *(Hawk flies)* ". The other is "木の葉が飛ぶ *(Leaves of trees scatter)* ". Figure 4.6 shows how to use the above procedure to analyze the Japanese sentence "雀が飛ぶ *(Sparrow flies)* ".

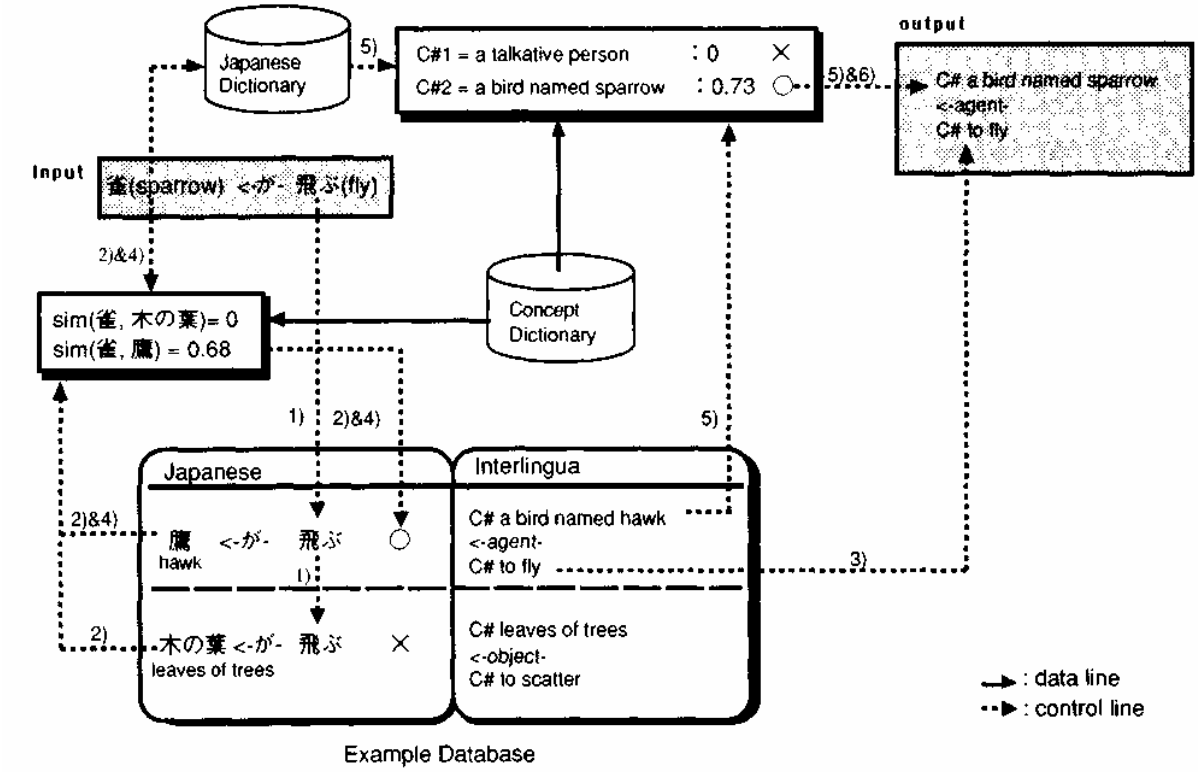


Figure 4.6: Semantic analysis based on Examples

Experiments using the semantic analysis system will be given in chapter 6.

It is possible to compare it directly with corpus data stored in the Example Database when input is a phrase or a simple sentence. On the other hand, in the case of a complex sentence, it will have to be broken down into a group of simple sentences and phrases. The semantic structure of the complex sentence is constructed based on the semantic structure of the simple sentence and phrases which are are contained in it.

This paper has described the EDR large scale mono-lingual corpus and an experimental Japanese semantic analysis system using the EDR corpora. With the help of EDR corpora and the

similarity calculation using the EDR Electronic Dictionary, it was possible to avoid writing complicated rules for semantic analysis.

5 The English Semantic Generation

This section describes English semantic generation. The objectives of the generation were restricted to content words and prepositions and other function words (articles, reflection postfixes, etc) were neglected in order to simplify the situation,

Below, at first examples used in the generation, next the similarities used by the generator and finally, the semantic generation using these similarities will be explained.

5.1 Examples Created from EDR English Corpus

The same as examples created in Japanese semantic analyzer, examples used in the English generation are pairs of an interlingua and a syntactic tree. Examples are created by dividing sentences in the EDR English corpus into simple sentences and phrases. Examples are created automatically by the translator. Figure 5.1 shows the created examples from the corpus data of Figure 2.4. Since restrictions of syntax are stronger in English than in Japanese, nodes that exist right above the head nodes of simple sentences or phrases in the syntactic tree of the example were added. (These added nodes are referred to as "dominator" below) In Figure 5.1, nodes circled are head nodes of simple sentences or phrases and nodes right above them are dominators.("<3cee21>" and "lik" in example 2, "<3cf458>" and "lik" in example 3 and "<0db4d>" and "book" in example 4 are dominators) It is assumed that a dominator of an interlingua and a dominator in a syntactic tree should correspond each other. " generation, examples are accessed only through head nodes of interlingua.

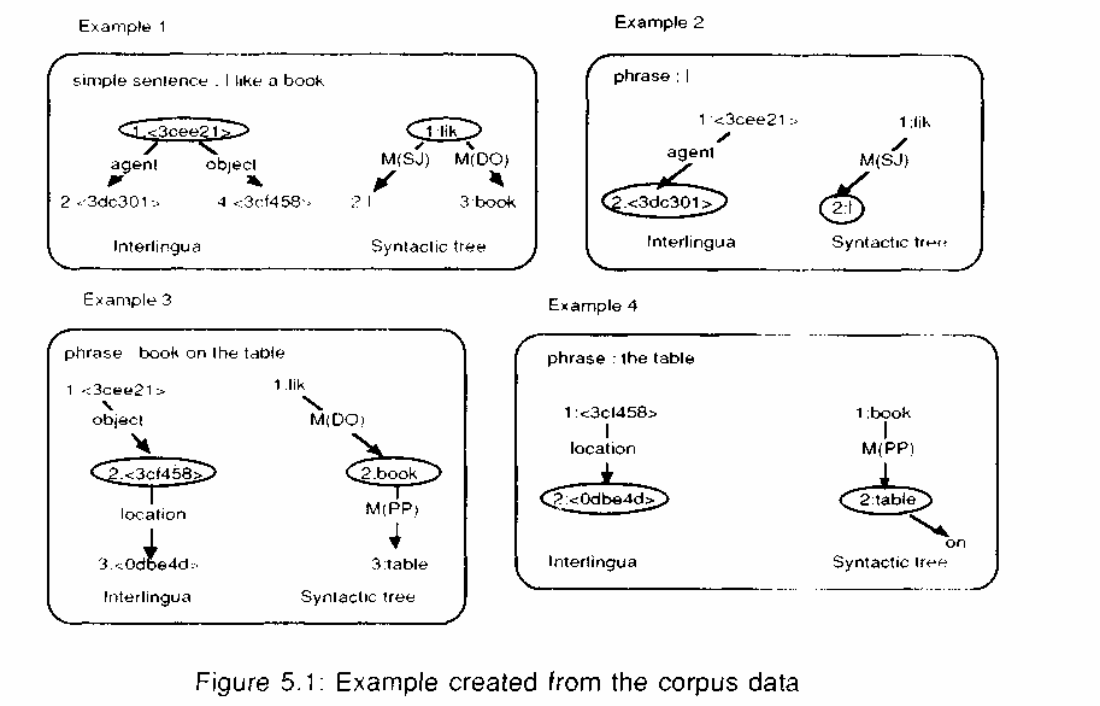


Figure 5.1: Example created from the corpus data

5.2 The Similarity Calculation

The similarities used in the generation differ from those used in the Japanese analyzer, since the roles of syntactic trees and interlingua are different here.

5.2.1 The Similarity between Concepts

The similarity between concepts is defined by word sets corresponding to concepts. The similarity between concepts is used to define the similarity between interlinguae.

Figure 5.2 shows correspondences between concepts and words. Intuitively the similarity between word set 1 and word set 2 is larger than the similarity between word set 1 and word set 3 or the similarity between word set 2 and word set 3.

In the generation system, such word sets corresponding to a concept were classified statically by line-ups of POS of word sets without repetitions. Figure 5.3 shows a part of clusters of word sets.

Word set 1 and word set 2 belongs to EN1_EN2 (not showed in Figure 5.2) and word set 3 belongs to EAJ_ED5.

Then the similarity between concepts are defined as follows:

The similarity between concepts C and C'

$= \sigma c(C,C')$

= "same" (=1) if C and C' are the same concepts,
"very similar" if C and C' belongs to the same cluster and are sisters in concept hierarchy, (5.1)
"similar" if C and C' belong to the same cluster.
"not similar" (=0) if C and C' do not belong to the same cluster.

Here quoted words mean some values with the order : "same" > "very similar" > "similar" > "not similar".

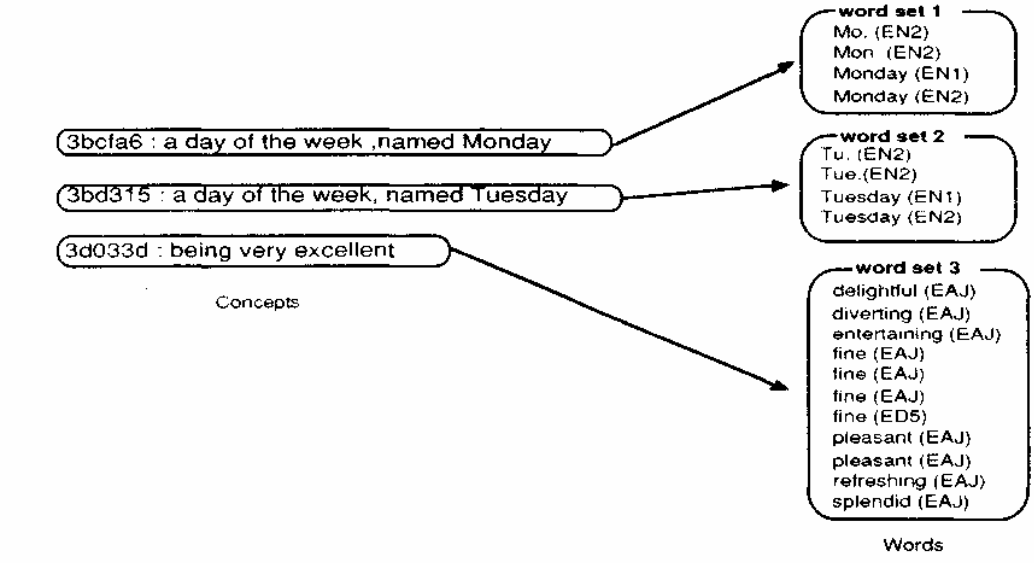


Figure 5.2: The correspondence between concepts and words

cluster number	line-ups of POS
1	EAJ
2	EAJ_ED3
3	EAJ_ED3_ED5
4	EAJ_ED3_ED5_EN1
5	EAJ_ED3_ED5_EN1_EVE
6	EAJ_ED3_EN1
7	EAJ_ED3_EVE
8	EAJ_ED5
9	EAJ_ED5_EIT
11	EAJ_ED5_EIT_EN1_EVE
12	EAJ_ED5_EN1
13	EAJ_ED5_EN1_EP4
14	EAJ_ED5_EN1_EVE

EVE=verb EAJ=adjective
EN1=common noun EN2=proper noun
EP4=pronoun ED3,ED5=adverb
EIT=interjection

Figure 5.3: A part of clusters of word sets used in the generation system

The similarities between concepts in Figure 5.2 are σc(<3bcfa6>,<3bd315>) = 1 and σc(<3bcfa6>,<3d033d>) = σc(<3bd315>,<3d033d>) = 0.

5 2.2 The Similarity between Words

The similarity between words is defined by their grammatical information. This similarity will be directly used to select words and to define the similarity between syntactic trees.

The similarity between words is defined as follows.

The similarity between words W and W'
= σw(W,W')
= "same" (=1) if W and W' have same spellings and POSs(Parts Of Speech).
"very similar" if W and W' have same POSs but have different spellings. (5.2)
"similar" if W and W' have replaceable POSs.
"not similar" (=0) if W and W' have different spellings and POSs.

Here quoted words mean some values with the order: "same word" > "similar words" > "not similar".

5.2.3 The Similarity between Simple Sentences or Phrases

Two similarities will be defined here : the similarity between interlinguae and the similarity between syntactic trees. The former is used to retrieve similar examples from the example database and the latter is used to evaluate generated syntactic trees to select the best one.

The similarity between interlinguae is the summation of similarities between concepts with some weights. The abstract definition of the similarity between interlinguae is as follows:

The similarity between interlinguae I and I'

$= \sigma i(I, I')$

$= \Sigma(\alpha i X$(the similarity between concepts with a relation i)) (5.3)

i varies in {all concept relations and head nodes and dominators}

Figure 5.4 shows weights of the similarities between concepts. Weights are given by concept relations that connect concepts to the head nodes of the interlingua. It is expected this similarity is used only in the generation, weights are decided by surface relations in the syntactic tree of the example . A head node of interlingua is an exception. Each concept relation should appear only once in an interlingua simple sentence or phrase. The relations between dominators and head nodes are neglected.

The ai was decided by using the detail relations as follows:

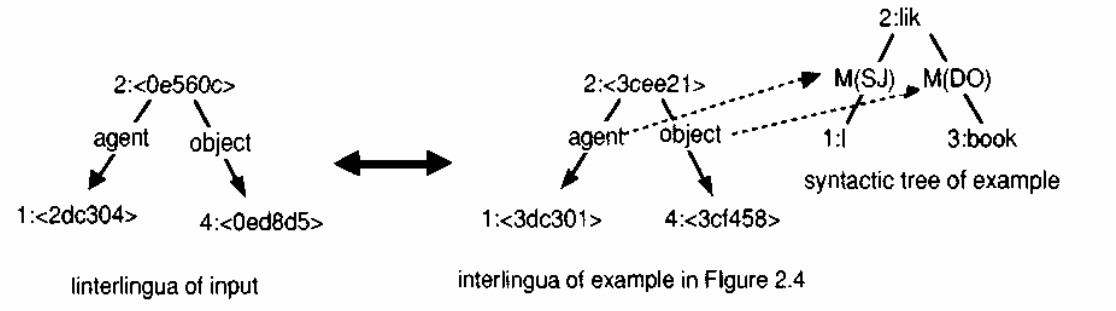
αhead = αdominator=" very large" (= 1)

αi = "large" if the surface relation corresponding to i in the syntactic tree is obligatory. "small" if the surface relation corresponding to i in the syntactic tree is optional.

(5.4)

In Figure 5.4, the agent relation and the object relation in interlingua are obligatory relations in the syntactical tree. So αagent and αobject are "large" by (5.4). ahead is always "very large".

The quoted words mean some values with the order: "very large" > "large" > "small".



α head ="very large" applied to concepts numbered 2 since concept head node
α agent = "large" applied to concepts numbered 1 since the surface relation corresponding agent is SJ
α object = "large" applied to concepts numbered 3 since the surface relation corresponding object is DO

Figure 5.4: Weights of the similarities between concepts of an interlingua

We add additional rules to check the agreements of arcs of the interlingua as follows :

1) The similarity between interlinguae = 0.

if the concept relation in the example corresponds to the obligatory relation (SJ,DO,IO,SC,OC) in the syntactic tree and either the concept in the input or in the example in same concept relation does not exist.

2) The similarity between concepts in this concept relation = 0,

if the concept relation in the example corresponds to the optional relation (PP.ADV.ADJ) in the syntactic tree and either the concept in the input or in the example in same concept relation does not exist. (5.5)

The definition of similarity between syntactic trees is as follows:

The similarity between syntactic trees T and T'

$$= \sigma t(T.T')$$

$$= \Sigma(\beta iX(\text{the similarity between words in surface relation i})) \quad (5.6)$$

i belongs to {all surface relations and head nodes and dominator}

Here, βi means the weight for the surface relation i.

We give the weights as follows:

$\beta head = \beta dominator =$ "very large" (=1)

$\beta SJ = \beta SJ = \beta DO = \beta IO = \beta SC = \beta 0C$ = "large" (in the case of obligatory relation) (5.7)

$\beta PP = \beta ADV = \beta ADJ$ = "small" (in the case of optional relation)

The words in quotations above have the values "very large" > "large" > "small" > 0. The relations between dominators and head nodes are neglected.

5.3 The Semantic Generation Based on Examples

When an interlingua is inputted, the generator generates words and surface relations. Figure 5.5 shows the order the generator deals with nodes. Each rectangle is a unit the generator deals with at one time, numbered in the order of processing.

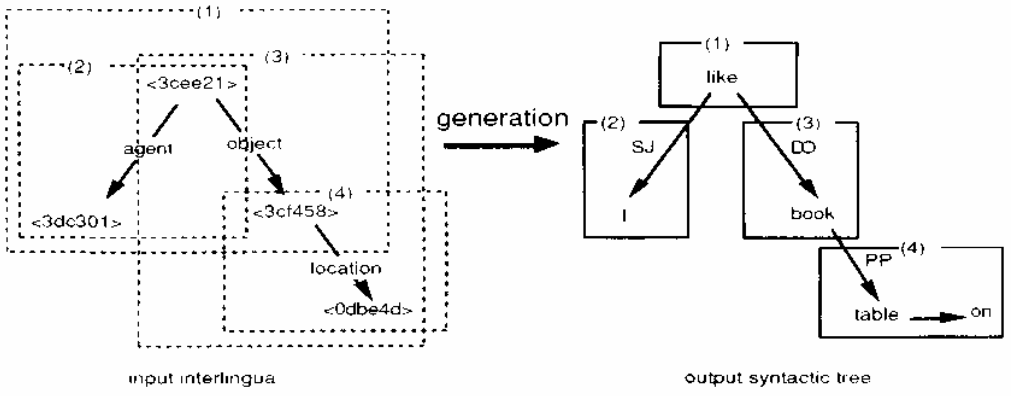


Figure 5.5: The order the generator deals with nodes

The Figure 5.6 shows the flow of the English semantic generation of simple sentences or phrases. The generation is accomplished as follows :

Step 1 : Get interlingua of a simple sentence or a phrase.
(Input interlingua below means this simple sentence or phrase)
Step 2 : Search the example database and retrieves some similar examples by using the similarity between interlinguae.
Step 3: Retrieve words corresponding to concepts in input interlingua from the word dictionary.
Step 4: Select words by using retrieved example and retrieved words by using the similarity between words. Make a syntactic tree with the selected words and surface relations of the example.
Step 5: Calculate the similarity between the syntactic tree of the example and the generated syntactic tree.
Step 6 : Repeat step 4 and step 5 by using other examples until the similarity between syntactic trees or the number of selected examples goes beyond some values.
Step 7 : Make an output syntactic tree.

Process of dividing input interlingua and combining output syntactic trees are not included the explain above.

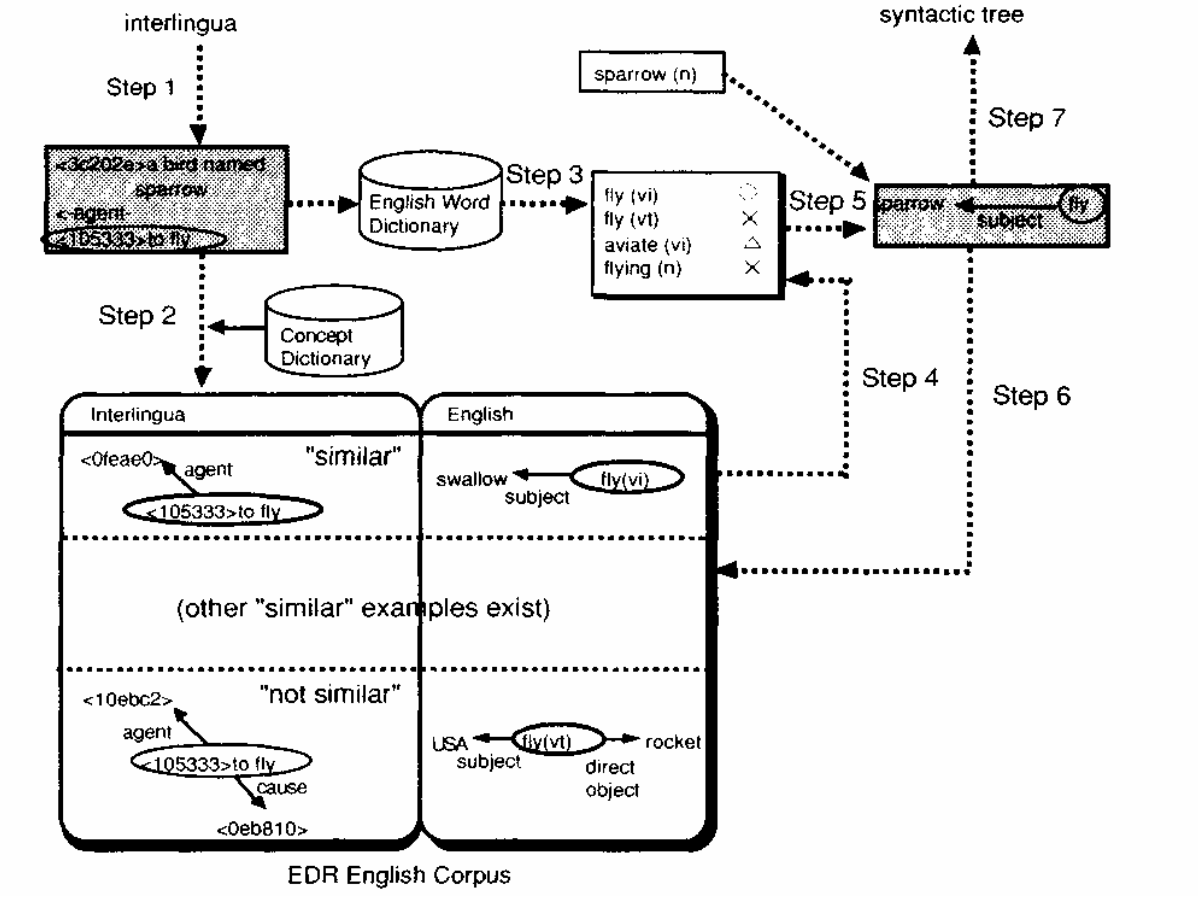


Figure 5.6: The flow of the English semantic generation of simple sentences or phrasesADJ

6.Experiments

6.1 An Experiment in Japanese Semantic Analysis

The number of records in Example Database is 68,830 that come from 3,035 sentences in EDR corpus base. The experiment is carried out for two purposes, one is to test the ability to use Examples in Example Database, the other is to test the speed to analyze sentences.

1)The ability to use examples in Example Database,

Various kinds of testing sentences that are similar to Examples in Example Database are made for this test. The ability to use Example in Example Database are tested by analyzing these testing sentences. The following are some related Examples in our Example Database

Example 1: 父は(my father) 電源開発の(of generating electric power) 仕事(job)のため(for) 山奥の(in the heart of mountain) 任地に(to the post) 赴いた(repair)。
Example 2: 母は(my mother) 急な(urgent) 用事で(on business) 出かけた(go out)。
Example 3: 山の(of mountain) 影が(image) 湖に(to the lake) 映っています(reflect)。
Example 4: 美しい少女に(to beautiful girl) 成長していた(grow up)。

Test 1: deleting phrases and replacing phrases

To Example 1, by

deleting the phrase 電源開発の仕事のため(for the job of generating electric power)
replacing the phrase 父は(father) to 姉は(my sister)
replacing the phrase 山奥の任地に(to the post in the heart of mountain) to 研究所に(to laboratory)

the testing sentence are made as follows:

姉は(my elder sister)研究所に(to laboratory)赴いた(go)。

as the result of semantic analysis, the system output interlingua

<3cf9b6> -agent-> <0e351f>
<3cf9b6> -goal-> <3cf44d>

Here:

<3cf9b6>: to go to the place one is aiming for
<0e351f>: a woman who is an elder sister in the same family
<3cf44d>: a place where a certain kind of work is done

Test 2: deleting phrases and adding phrases of other similar examples

To Example 1, by

deleting the phrase 電源開発の仕事のため(for the job of generating electric power)
replacing the phrase 父は(my father) to 兄は(my brother)
replacing the phrase 山奥の任地に(to the post in the heart of mountain) to 作業場に(to the workshop)
adding 用事で(on business) in Example 2

the testing sentence are made as follows :

兄は(my brother) 用事で(on business) 作業場に(to workshop) 出かけた(go out)。

as the result of semantac analysis, the system output interlingua

<3cf9b6> -agent-> <0e34eb>
<3cf9b6> -cause-> <3cec3d>
<3cf9b6> -goal-> <3cf44d>

Here:

<3cf9b6>: to go to the place one is aiming for
<0e34eb>: a person who is an elder brother
<3cec3d>: something that is to be done by someone
<3cf44d>: a place where a certain kind of work is done

Test 3: replacing phrase and modifier of the phrase, adding modifier to the phrase

To the sentences in Example3 and Example4 :

replacing the phrase 山の影(image of mountain) to 少女の姿(figure of girl)
adding modifier 美しい(beautiful) to 少女(girl)

the testing sentence are made as follows :

美しい少女の(of beautiful girl) 姿が(figure) 川に(to river) 映っています(reflect)。

as the result of semantic analysis, the system output interlingua

<0e5ed4> -object-> <3cf720>
<0e5ed4> -goal-> <0ea9dd>
<3cf720> -modify-> <0e5a95>
<1e84c3> -object-> <0e5a95>

Here:

<0e5ed4>: a shape or a colour to be reflected in a mirror or water surface
<3cf720>: the appearance of a person or thing
<0ea9dd>: a waterway where natural water gathers and flows
<0e5a95>: a young woman
<le84c3>: being beautiful

2) the speed of analysis of a sentence,

The system was tested with more than 100 sentences. The average time to analyze one sentence is approximate 5 second. The average time to retrieve one related example is approximate 0.45 second.

The retrieval speed becomes a serious problem when the Example Database is huge. For the purpose of preventing this problem, instead of searching all examples in the Example Database to find similar examples, in this system similar examples are retrieved by making use of a retrieving key. We just seek some definite records of the Example Database rather than all of them.

6.2 The Experiments of the English Semantic Generation

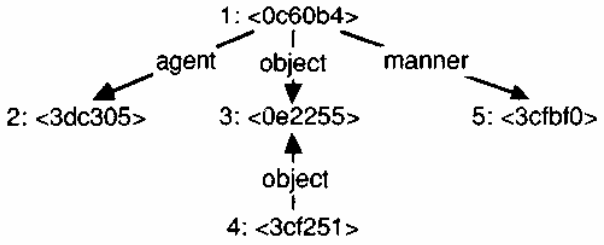
The number of records in Example Database is 35,434 that come from 1,744 sentences in EDR corpus base. For experiments, the similarities are simplified and some procedures are introduced to control the search. In the similarity between interlinguae, only « head is meaningful and other weights is set to 0, though the generator checks if configurations of conceptual relations are similar by additional rules.

The retrieval is controlled by the procedure. At first, examples with the "same" head node of an interlingua are searched. At next, examples with "very similar" nodes are searched and finally examples with "similar" nodes are searched.

Qualities and speed depend on the number of examples. The generator restricts the range of searching by two parameters: 1) the border value of the similarity between the generated syntactic tree and that of the example and 2) the maximum number of units (simple sentences or phrases) to be generated. If the similarity goes beyond parameter 1 or the number of generated sentences goes beyond parameter 2, the generator quit searching more examples and output the syntactic tree with the largest similarity.

Below is a sample of the execution of the generator. The first parameter is 0.5 and the second is 2. As for the similarity between words, "same" = 1.0, "very similar" = 0.9, "similar" = 0.8 and "not similar" = 0. As for the weights for the similarity between simple sentences or phrases of syntactic trees, "very large" = 1.0, "large" = 0.9 and "small" = 0.8.

Input interlingua is :



<3dc305> : c#she

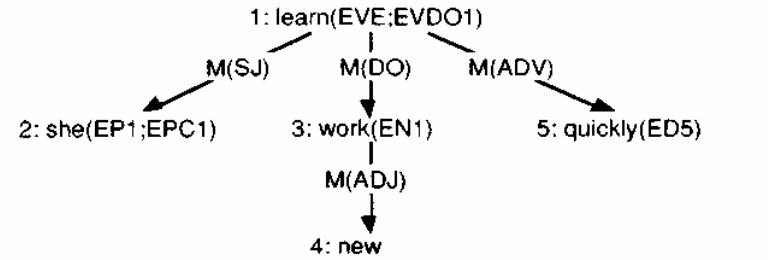
<0c60b4> : to gain knowledge of (a subject) or skill in (an art, trade, or other specialty)

<0e2255> : activity which uses effort, especially with a special purpose, not for amusement

<3cf251> : a condition of being quick

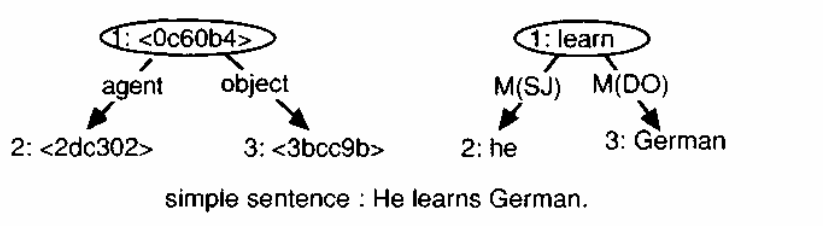
<3cfbfO> : new; unlike others of the same type

Output syntactic tree is :



sentence : She learns new work quickly.

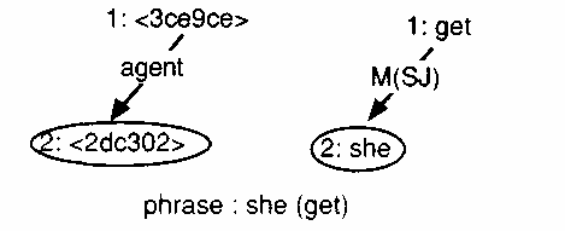
1 : "learn" was generated by using the example:



simple sentence : He learns German.

$\sigma t = \sigma w("","") \times \beta dominator + \sigma w("learn","learn") \times \beta head + \sigma w("she","he") \times \beta SJ$
$+ \sigma w("work","German") \times \beta DO$
$= 1.0 \times 1.0 + 1.0 \times 1.0 + 1.0 \times 1.0 + 0.9 \times 0.9 + 0.9 \times 0.9 = 3.62$.
(If dominators don't exist, the similarity between dominators is assume to be "very large")

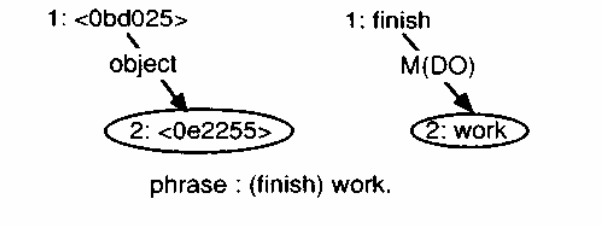
2 : "she" was generated by using the example :



phrase : she (get)

$\sigma t = \sigma w("learn","get") \times \beta dominator + \sigma w("she","she") \times \sigma head$
$= 0.9 \times 1.0 + 1.0 \times 0.9 = 1.80.$

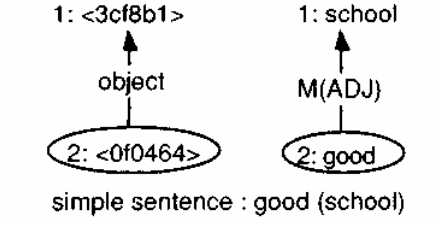
3 : "work" was generated by using the example:



phrase : (finish) work.

$\sigma t = \sigma w("learn","finish") \times \beta dominator + \sigma w("work","work") \times \beta head$
$= 0.9 \times 1.0 + 1.0 \times 1.0 = 1.90.$

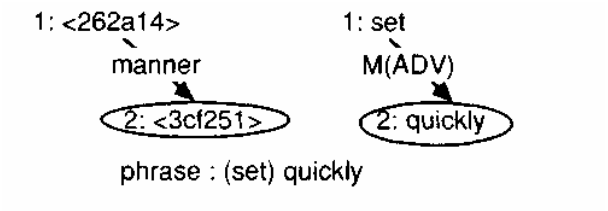
4 : "good" was generated by using the example:



simple sentence : good (school)

σt. =σw("work","school") X βdominator +σw("new","good") X βhead
= 0.9 X 1.0 + 0.9 X 1.0 = 1.80.

5 : "quickly" was generated by using the example:



σt = σw("learn","set") X βdominator + σw("quickly","quickly") X βhead
= 0.9 X 1.0 + 1.0 X 1.0 = 1.90 .

The parentheses in simple sentences and phrases mean dominators.

The generator used more examples, but shown here is the example that has the largest similarity of syntactic trees.

It takes about 0.25 seconds to search one example with "same" head nodes of an interlingua and generate a part of a syntactic tree with it. If such an example doesn't exist, it takes more time.

7. Concluding Remarks

This paper has described the prototype of an example-based machine translation system and experiments, mainly from the point of the view of how to use concepts and corpora. With the help of the EDR corpora and the similarity calculation using the EDR Electronic Dictionary, we got away from writing complicated rules for semantic analysis was avoided. Expansion of dictionaries (including corpora) will solve some problems of example-based MT. So, at this point, there exists a good prospect for success of the example-based machine translations.

Emphasis is given to the combination of example-based MT and interlingua(or mono-lingual corpus). Some advantage of this combination was found, the easiness of creating corpora, for example. This method may involve many theoretical problems [9]. So, more research for this combination is necessary. [14]

Reference

[1] Cui, J.; Komatsu, E. and Yasuhara, H. : "A Calculation of Similarity between Words Using EDR Electronic Dictionary", Reprint of IPSJ, Vol.93, No.1 (January 1993) (in Japanese)

[2] EDR: Concept Dictionary, TR-027(1990)

[3] EDR: EDR Electronic Dictionary Specification Guide, TR-041(1993)

[4] EDR: English Word Concept Dictionary, TR-026(1990)

[5] EDR: Japanese Word Dictionary, TR-025(1990)

[6] Furuse, O. and Iida, H. : "An Example-Based Method for Transfer-Driven Machine Translation", Fourth International Conference on Theoretical and Methodological Issues

in Machine Translation, pp139-150 (June 1992)

[7] Komatsu, E.; Cui, J. and Yasuhara, H. : "English generation from EDR concept relation representation" Proc. of IPSJ 45th, pp323-324 (August 1992)

[8] Nagao, M. : "A Framework of A Mechanical Translation between Japanese and English by Analogy Principle, Artificial and Human Intelligence (A. Elithorn and R. Banerji, editors) Elsevier Science Publishers, B.V. (1984)

[9] Sadler, V. : Working with Analogical Semantics: Disambiguation Techniques in DLT, Foris Publications, Dordrecht Holland, (1989)

[10] Sato, S. : "Example-Based Translation Approach" Proc. of International Workshop on Fundamental Research for the Future Generation of Natural Language Processing, ATR Interpreting Telephony Research Laboratories, pp. 1-16 (1991)

[11] Sumita, E. and Iida, H. : "Example-Based Transfer of Japanese Adnominal Particles into English", IEICE TRANS. INF. & SYST., VOL. E75-D, NO.4 (July 1992)

[12] Tanaka.H. : "Foundation of Natural Language Understanding", Vol.2-3, pp.9-154

[13] Uchida, H. and Zhu M.:" An Interlingua for Multilingual Machine Translation, 89-NL-72-9, Information Processing Society of Japan (1989)

[14] Yasuhara, H. : "An Example-Based Multilingual MT System in a Conceptual Language", Proc. of MT Summit 4, (forthc. July 1993)