ENGLISH GENERATION FROM INTERLINGUA
BY EXAMPLE-BASED METHOD

Eiji Komatsu *,   Jin Cui **,   Hiroshi Yasuhara **

(*) Oki Electric Industry Co. Ltd. Multimedia Laboratory
11-22, Shibaura 4-Chome, Minato-ku, Tokyo 108 Japan
e-mail : komatsu @okilab.oki.co.jp

(**) Japan Electronic Dictionary Research Institute Ltd. (EDR) 6th Laboratory
Mita-Kokusai-Bldg. 4-28, Mita 1-Chome, Minato-ku, Tokyo 108 Japan
e-mail : sai@edr6r.edr.co.jp, yasuhara@edr6r.edr.co.jp

Abstract

This paper describes the experiment of the English generation from interlingua by the example-based method. The generator is implemented by using English Word Dictionary and Concept Dictionary developed in EDR. How to construct examples and how to define the similarities are main problems. The results of experiments are shown.

1. Introduction

This paper describes the generator that is originally implemented to correct and evaluate English Word Dictionary and Concept Dictionary being developed in EDR (EDR,1993). To evaluate Concept Dictionary, as the first strategy, interlingua method was introduced. As the number of concepts is very large and they are elements of complex hierarchy, it is difficult to make rules and on the other hand the example-based method was expected to be more effective than the rule-based method. So, as the second strategy, the example-based method was also introduced.

The example-based method is usually used in MT by the transfer method (Nagao, 1984; Sato, 1991; Sumita, 1992), though one by Sadler (1989) is by the interlingua method. In this generator, the example-based method co-exists with the interlingua method because of above reasons, but the combination of the example-based method and the interlingua method is not important, because from another point of view, the generation from interlingua is recognized as a translation from one language i.e. interlingua to another i.e. English and the generation from interlingua can be seen similar as translations in above MT systems. So in this experiment, how to apply the example-based method to various natural language processing and for which parts the method are suitable are the main interests. For this purpose, the generator is designed to execute the generation with maximum usage of the example-based method.

In this experiment, the coverage of the generation is not complete, that is, some elements such as articles and conjunctions are not generated.

Below, section 2 describes the input and output of the generator, section 3, examples used in this system, section 4, the similarities used to retrieve examples and to select words, section 5, the generation algorithm, section 6, the experiments for verb selections and section 7, the conclusion.

The examples, similarities and the generation algorithm are decided a priori then modified in response to the output of the generator.

To avoid confusions, the word "example" is used only

_____

(*) This work has been done when the author was in EDR.

to mean example data of the example-based method. And the terms "interlingua" and "syntactic tree" are used to mean sets, elements and fragments of elements.

2. Input and Output

The generator translates an interlingua to a syntactic tree. Fig.2.1 shows a sample of input interlinguae and Fig.2.2, a sample of output syntactic trees. Both samples correspond to the same sentence "My brother will take the medicine".

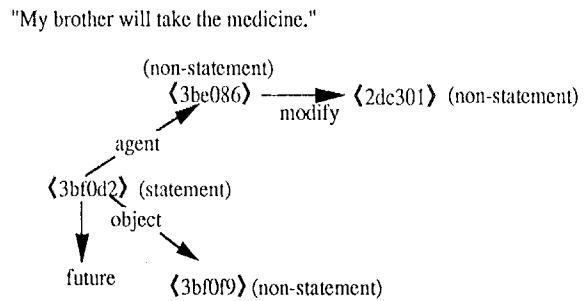"My brother will take the medicine."

Fig.2.1 Input Interlingua

Interlinguae consist of concepts, conceptual relations and attributes. Each concepts are classified as "statements" or "non-statements". Concepts are represented by concept identification numbers (To distinguish concepts easily by men, concept illustrations are also given). Interpretations of codes relating to interlinguae in this paper are shown in Table 2.1. In the table, as for concept identification numbers, concept illustrations are showed as interpretations of codes.
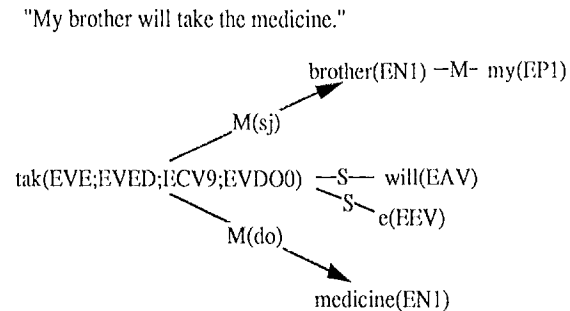
"My brother will take the medicine."

Fig.2.2 Output Syntactic Tree

Table 2.1 Codes for Interlinguae

Information	Code	Interpretation
Concept Identification Numbers	⟨3bf0d2⟩	to drink something
	⟨3be086⟩	brothers
	⟨0e351f⟩	sisters
	⟨2dc301⟩	c#I
	⟨2dc304⟩	c#he
	⟨3bf0f9⟩	a substance used on or in the body to treat a disease
	⟨3bdbf6⟩	a drilled liquor named whiskey
	⟨3bd862⟩	a drug or agent that reduces fever
	⟨3cee4f⟩	to obtain a thing which one wanted
	⟨3ceae3⟩	to become a certain condition
	⟨0fdc5f⟩	to accept others' opinions and wishes
	⟨0c98dc⟩	the first part of the day, from the time when the sun rises, usually until the time when the midday meal is eaten
Concept Relations	agent	Subject that brings about a voluntary action. Conscious and automated entities are such subjects. "Animals eat" ⟨eat⟩ — agent—▶ ⟨animals⟩
	object	Object affected by an action or change "Eat food." ⟨eat⟩—— object—▶ ⟨food⟩
	time	Time at which an event begins "Work until a se time" ⟨wake up⟩ —— time—▶ ⟨in time⟩
	modifier	Other relationships
	past	Viewpoint is in the past
	present	Viewpoint is in the present
	future	Viewpoint is in the future
	end	The end of an action or event
	already	Already occurred

Table 2.2 Codes for Syntactic Trees

Information	Code	Interpretation
Part-of_Speech	EN1	Common noun
	EP1	Personal pronoun
	EVE	Verb
	EAV	Auxuriary verb
	EEV	Verb ending
	EPR	Preposition
	EAR	Article
Grammatical Information	EVSTM	Uninflected part
	EVB	Infinitive
	EVED	Past tense
	EVEN	Past participle
	ECV9	Partially irregular inflections ("e" follws)
	EVDO0	Takes a direct object
	EVDO6	Takes a direct object (the direct object is to-infinitive)
Surface Relations	M(sj)	subject relation
	M(do)	direct object relation
	M(adj)	adjective modification
	M(obpp)	obligatory prepositional phrase
	S	relations between content words and functional words

Syntactic trees consist of words, part-of-speeches, grammatical information and syntactic relations.

The interpretations of codes relating to syntactic trees used in this paper are shown Table 2.2.

3. Examples

An example should be a pair of an interlingua and a syntactic tree. For the flexibility of usage of examples, interlinguae and syntactic trees in examples are divided into smaller parts that are small enough to use flexibly but have enough information for generations.

Fig.3.1 shows the common form of interlinguae and syntactic trees in examples (referred as "basic unit", below). An example is a pair of fragments in this form made from an interlingua and a syntactic tree.
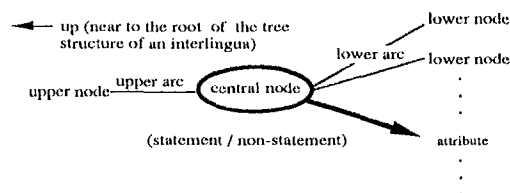

Fig.3.1 Basic Units

Fig.3.2 shows the linguistic resources used by the generator. As the results of trying to execute as many processes as possible by the example-based method, it became necessary for the generator to use two different kinds of examples (referred as "Basic Example Set " and "Example Set for Attribute", below).
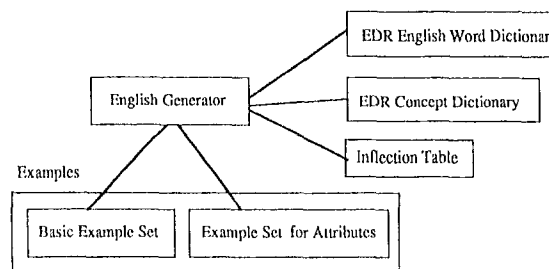

Fig.3.2 Linguistic Resources

Fig.3.3 shows examples in the Basic Example Set. Circled nodes are "central nodes" . Basic Example Set is supposed to be used for selecting content words for concepts. Functional words except prepositions and grammatical information for inflections are removed, since they are unnecessary for this purpose. In Fig.3.2, example (A) and (B) have no upper node and Example (C) and (D) have no lower node. Examples in this set are accessed by concepts in the central nodes of interlinguae; Example (A) and (B) are accessed by ⟨3bf0d2⟩ and (C), by ⟨3bf0f9⟩ and (D) by ⟨0c98dc⟩ . When several examples with the same key exist, by the similarity defined below, only one example is finally accepted.

Fig.3.4 shows examples in the Example Set for Attributes. This example set is supposed to be used for deciding inflection (i.e. selecting the word whose inflection corresponds to the attributes) and adding functional words for attributes. Content words in lower nodes are

removed, since the upper node influences to the inflection of the center word, but the lower nodes rarely don't. Functional words in lower nodes are added to the outputs. Concepts and spellings of words are also removed, since they can be decided by Basic Example Set and unnecessary here. Examples are accessed by combinations of attributes in interlinguae, some grammatical information of the upper node, those of central nodes and the surface relation of the upper arc; in Fig.3.4, Example (a) is accessed by (past, -, EVE; EVED, -), Example (b) by (end, already, -, EVE; EVEN, -), Example (c) by (present, -, EVE; EVSTM; ECV9, -), Example (d) by (present, , -, EVE; EVB, -), Example (e) by (future, -, EVE; EVSTM; ECV9,-) and Example (f) by (-, EVE; EVDOO, EN1, M(do)). Example (a), (b), (c), (d) and (e) have no upper node. Since examples in this set don't include concepts, examples are accessed deterministically and the similarity is not used.

4. Similarities

There are two major similarities in the example-based method. One is for the source language and used for selecting examples. Another is for the target language and used for creating outputs. In this generator, the former is the similarity between interlinguae (in the form of basic units) and the latter is the similarity between words. In the generator, the similarity is used only for Basic Example Set.
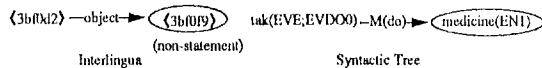
Example (A) : Brother takes the medicine in the morning.
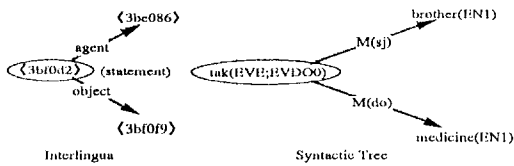
Example (B) : Sister drinks the whiskey.
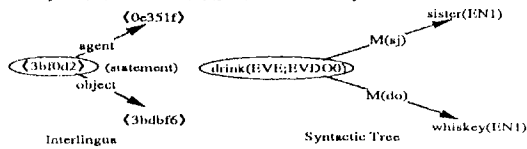
Example (C) : Brothers takes medicine in the morning.

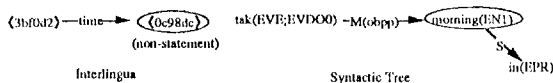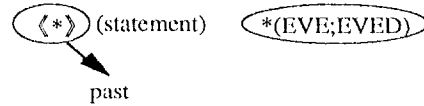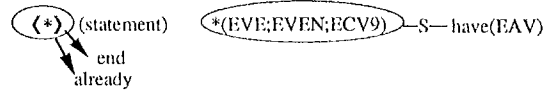Example (D) : Brothers takes medicine in the morning.

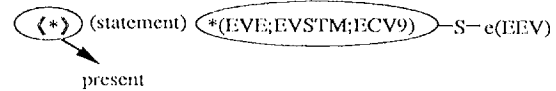Fig.3.3 Examples in Basic Example Set

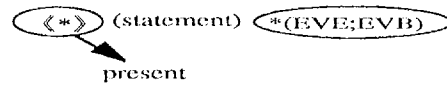Example (a) : *(EVE;EVED;EVDOO)
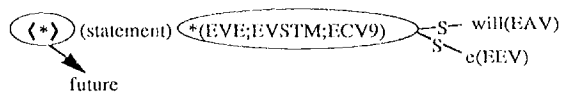

Example (b) : have *(EVE;EVEN;EVDOO)

Example (c) : *(EVE;EVSTM;ECV9;EVDOO) e

Example (d) : *(EVE;EVB;EVDOO)

Example (e) : will *(EVE;EVSTM;ECV9;EVDOO) e
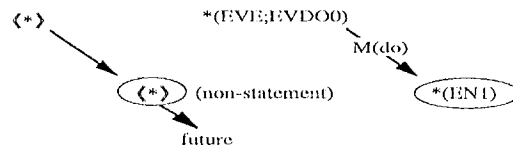

Example (f) : *(EVE;EVDOO) *(EN1)

Fig.3.4 Examples in Example Set for Attributes

The similarity between interlinguae is defined as follows;

$$Si(IL1,IL2) = (Sc(C1cent,C2cent) \times Kcent$$
$$+ \Sigma \, Sc(C1i,C2i) \times K(srel(i)) \times (k(num(R1 \cap R2) + 1)$$
$$i \in R1 \cap R2$$

IL1,IL2 : interlinguae
C1cent, C2cent : concepts in central nodes
Kcent : weight of similarity between central nodes
C1i, C2i : concepts in lower nodes with arc i
k(x) : weight of similarity between concepts in lower nodes, x is the number of elements in the interjunction
srel(i) : surface relation which corresponds to the concept relation i
R1,R2 : set of conceptual relations each for IL1, IL2
num(S) : the number of elements of set S

It is always assured in advance by the generator that 1) the word in the upper node of the input is already selected (if there is an upper node); 2) arcs of interlingua, which correspond to obligatory relations of the syntactic tree in the example, exist in the interjunction of R1 and R2; 3) upper arcs are same (if already decided); 4) part-of-speeches of words in upper nodes are same. Examples that don't satisfy these

four conditions are rejected before the similarity calculation.

The similarity between concepts used in the above similarity is defined as follows;

$$Sc(C1,C2) = \frac{\text{the number of common ancestors}}{\text{the number of ancestors of C1 + the number of ancestors of C2}}$$

Here, ancestors until three layers above are used. (Cui; 1993)

It is difficult to find the most similar interlingua in an example set to the input interlingua, because to find it , it is necessary to calculate all similarities between interlinguae in the example set and the input. To avoid this, in this generator, some constraints are given for access keys i.e. central nodes. For "statements" in interlingua, central nodes of examples should be same with that of the input and for "non-statements" in interlingua, central nodes of examples can be the same concepts or sister concepts in the concept hierarchy. By this constraints, the search of examples can be executed fast.

The similarity between words is defined as follow;

$$Sw(W1,W2) = \begin{cases} 1 \text{ if spelling part-of-speech and grammatical information} \\ \text{are same.} \\ k\ (0 < k < 1)\ \text{if part-of-speech and grammatical information} \\ \text{are same} \\ l\ (0 < l < 1)\ \text{if part-of-speech are same} \\ 0\ \text{if speling, part-of-speech and grammtical information are} \\ \text{all different} \end{cases}$$

k, l : some numbers

5. Generation Algorithm

The generator generates fragments of a syntactic tree and finally combines them into a syntactic tree.

The generation algorithm is as follows;

Step 1 : Sets the current central node at the root node of the input interlingua.

Step 2-1 : Cuts the basic unit for the current central node.

Step 2-2 : Extracts candidate English words for concepts of the central node and lower nodes of the current basic unit, from English Word Dictionary.

Step 3-1 : Retrieves an example from Basic Example Set.

Step 3-2 : Selects the same word (neglecting inflection) from the candidate word lists and checks if there is an example in Example Set for Attributes, whose attributes and words in the central node coincide with attributes in the current basic unit and the selected word.

Step 3-3 : If the word selection succeeded, accepts the example. Generates upper arc (if exists), lower arc (only for obligatory relations) central nodes and functional words for the central node, saves the results and similarity and calculates the similarity of interlingua between the input and the example. Prepositions are extracted from the basic example.

Step 3-4 : Repeat Step 3-2 to Step 3-3 until there remains no basic examples.

Step 3-5 : Selects one example that is accepted in Step 3-3 and the similarity is largest.

Step 3-6 : Puts the results.

Step 4 : Move the current central node in the input linterlingua in depth-first order.

Step 5 : Repeat Step 2-1 to Step 4 until the movement of the current central node ends or the word selection for a node fails.
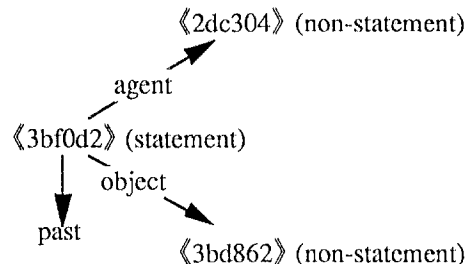

Figure 5.1 Inputted Interlingua

Suppose the interlingua such as Fig.5.1 is inputted and examples in Fig.3.3 are used as Basic Example Set and Fig.3.4 used as Example Set for Attributes.

The list of candidate words for ⟨3bf0d2⟩ is as follows;

tak(EVE;EVSTM;ECV9;EVDO0),
took(EVE;EVED;EVDO0),
taken(EVE;EVEN;EVDO0),
drink(EVE;EVB;EVDO0),
drank(EVE;EVED;EVDO0),
drunk(EVE;EVEN;EVDO0).

From Basic Example Set, Example (A) and (B) are retrieved, since central nodes are same.

By Example (A) and Example (a), took(EVE; EVED; EVDO0) is selected and by Example (B) and Example (a), drank(EVE; EVED; EVDO0) is selected.

As similarity between the input and Example (A) is larger than that between the input and Example (B), "took" is selected. This is because similarity between ⟨3bd862⟩ but ⟨3bf0f9⟩ is 0.876535 and one between ⟨3bd862⟩ and ⟨3bdbf6⟩ is 0.

6. Experiments for Verb Selections

This chapter describes experiments to evaluate examples, similarities and the generation algorithm. Experiments for verb selections are executed.

The generator selects one word from candidate word list retrieved from EDR English Dictionary.

The experiments are done by Jack-knife test method (Sumita; 1992) ; 1) Specify a concept; 2) Collect examples that include a word in candidate word list whose meaning is same with the specified concept ; 3) Remove one example from example sets; 4) Make the input interlingua from the removed example; 5) Generate a sentence from this interlingua by using remained examples; 6) Compare the original word and the generated word for the verb; 7) Repeat 3) - 6) by removing each example in turn.

Below the results of three experiments (Experiment 1, Experiment 2, Experiment 3) are shown.

Table 6.1 shows specified concepts for experiments and candidate word lists for the concepts. As for Experiment 1 and Experiment 2, words that have no examples is omitted from candidate word lists, since they won't never be selected. Fig.6.1, Fig.6.2 and Fig.6.3 show examples and generated sentences for Experiment 1, Experiment 2 and Experiment 3 each. Examples in Fig.6.1 and Fig.6.2 are extracted from EDR English Corpus and examples in Fig.6.3 are extracted from a published printed English-Japanese dictionary, though some modifications (Tenses, aspects ,

modals are all same. Subjects are same if possible) are done.

Sentences in the left hand sides of arrows are original sentences and those in the right hand side are generated sentences (In generated sentences, only verbs are generated words and others are copied from original sentences). Underlined words are words for the specified concepts. For sentences with a circle at the head of left hand sides, the generator selects same words with those in the original sentences. Sentences without circles include both right and wrong results.

In interlingua method, roughly speaking, all words corresponding to a concept are basically right as the generated word if it is grammatically consistent. So the evaluation of the experiments is delicate.

The rates of coincides between original verbs and generated verbs are 85% (Experiment 1), 13% (Experiment 2) and 16% (Experiment 3). Since some sentences without coincides can be also right, the real rates of success are lager than above numbers.

7. Conclusions

The English generation by the example-based method is described. For experiments of verb selections, the effectiveness of the method is different for verbs to be generated. (In experiment 3, for "confirm" and "endorse" the success rate is high). It also depends on concepts and the number of candidate words.

Since examples are made automatically from large scale corpus and to make examples is easier than to make rules, the effort to design the generator became little. By removing redundant basic units, the efficiency of examples is not serious.

In this paper, only the experiments for verb selections are shown. But the strategies that the generator uses should vary in response to the categories of words to be generated. For example, to generate prepositions the semantic is more important, but to generate other functional words the syntax is more important. For verb selections, both are necessary. These strategies are also remained problems.

Table 6.1 Concepts and Word List

Expriments	Specified Concept	Candidate Word List
Experiment 1	《3cee4f》	acjoev(e)(EVE) get(EVE) tak[E] (EVE) (others are omitted)
Experiment 2	《3ceae3》	get (EVE) grew (EVE) fall (EVE) (others are omitted)
Experiment 3	《0fdc5f》	accept (EVE) acknowledg(e) (EVE) admit (EVE) allow (EVE) answer (EVE) approv(e) (EVE) confirm (EVE) endors(e) (EVE) grant (EVE) receiv(e) (EVE) ratif(y) (EVE) recogniz(e) (EVE) respond (EVE) homologat(e) (EVE)

ex. 01 :   He had achieved a certain transquility.
    → He had got a certain transquility.
ex. 02 :   ○You have got our keys.
    → You have got our keys.
ex. 03 :   ○He quietly got a broom.
    → He quietly got a broom.
ex. 04 :   ○He got the menus.
    → He got the menus.
ex. 05 :   ○In the storm I took shelter under a tree.
    → In the storm I took shelter under a tree.
ex. 06 :   ○He takes dangerous drugs.
    → He takes dangerous drugs.
ex. 07 :   ○The people took our old house.
    → The people took our old house.

Fig.6.1 Examples and Results of Experiment 1

ex. 01 :   Diamonds come expensive.
    → Diamonds become expensive.
ex. 02 :   You grow older.
    → You become older.
ex. 03 :   A thing was becoming increasingly sure.
    → A thing was getting increasingly sure.
ex. 04 :   Environment becomes individualized.
    → Environment grows individualized.
ex. 05 :   A man gets old anyhow.
    → A man becomes old anyhow.
ex. 06 :   These letters became the center of my existence.
    → These letters went the center of my existence.
ex. 07 :   Almost unbearable my fantasies become.
    → Almost unbearable my fantasies go.
ex. 08 :   Something had gone wrong.
    → Something had fallen wrong.
ex. 09 :   We had become good friends during my stay at the hospital.
    → We had grown good friends during my stay at the hospital.
ex. 10 :   You're the kind to go violent.
    → You're the kind to become violent.
ex. 11 :   ○Her eyes became bright.
    → Her eyes became bright.
ex. 12 :   Eventually it become a movie.
    → Eventually it got a movie.
ex. 13 :   After a while the signal became a buzz.
    → After a while the signal went a buzz.
ex. 14 :   It was getting light.
    → It was becoming light.
ex. 15 :   He fell silent, as yesterday.
    → He went silent, as yesterday.
ex. 16 :   After a few jokes his speech became serious.
    → After a few jokes his speech went serious.
ex. 17 :   You'll get even fatter.
    → You'll grew even fatter.
ex. 18 :   She became stout.
    → She grew stout.
ex. 19 :   The fish has gone bad.
    → The fish has become bad.
ex. 20 :   ○He suddenly became wealthy.
    → He suddenly became wealthy.
ex. 21 :   She became impatient.
    → She went impatient.
ex. 22 ; ○He became a priest.
    → He became a priest.

Fig.6.2 Examples and Results of Experiment 2

ex. 01:   I accept an invitation.
  → I allow an invitation.
ex. 02 :   I accept an offer.
  → I receive an offer.
ex. 03 :   I acknowledge a defeat.
  → I accept a defeat.
ex. 04 :   I acknowledge his right.
  → I grant his right.
ex. 05 :   I acknowledge the truth of an argument.
  → I recognize the truth of an argument.
ex. 06 :   I admit a claim.
  → I allow a claim.
ex. 07 :   I admit defeat.
  → I acknowledge defeat.
ex. 08 :   I admit my guilt.
  → I acknowledge my guilt.
ex. 09 :   I will admit no objection.
  → I will accept no objection.
ex. 10 :   I allow a claim.
  → I recognize a claim.
ex. 11 :   I allow your argument.
  → I confirm your argument.
ex. 12 :   I answer his wish.
  → I receive his wish.
ex. 13 :   I approve a bill.
  → I accept a bill.
ex. 14 :   I approve a resolution.
  → I confirm a resolution.
ex. 15 :   I approve accounts.
  → I accept accounts.
ex. 16 : ○I confirm a treaty.
  → I confirm a treaty.
ex. 17 : ○I confirm an appointment.
  → I confirm an appointment.
ex. 18 :   I confirm a verbal promise.
  → I approve a verbal promise.
ex. 19 :   I confirm a telegraphic order.
  → I answer a telegraphic order.
ex. 20 :   I confirm possession to him.
  → I acknowledge possession to him.
ex. 21 :   I confirm a functionary in his new office.
  → I accept a functionary in his new office.
ex. 22 : ○I endorse his opinion.
  → I endorse his opinion.
ex. 23 : ○I endorse a policy.
  → I endorse a policy.
ex. 24 :   I grant a request.
  → I acknowledge a request.
ex. 25 :   The king granted the old woman her wish.
  → The king answered the old woman her wish.
ex. 26 :   Japan receive a treaty.
  → Japan ratifies a treaty.
ex. 27 : ○Parliament ratified the agreement.
  → Parliament ratified the agreement.
ex. 28 :   I receive a proposal.
  → I accept a proposal.
ex. 29 :   I receive an offer.
  → I accept an offer.
ex. 30 :   I receive a petition.
  → I answer a petition.
ex. 31 : ○Priest receives his confession.
  → Priest receives his confession.
ex. 32 :   Priest receives his oath.
  → Priest ratifies his oath.
ex. 33 :   I recognize a claim as justified.
  → I allow a claim as justified.

Fig.6.3 Examples and Results of Experiment 3

ex. 34 :   Japan recognizes the independence
  of a new state.
  → Japan acknowledges the independence of ...
ex. 35 :   He responds quickly to the appeal
  for subscriptions.
  → He approves quickly to the appeal for ...

Fig.6.3 Examples and Results of Experiment 3 (remainder)

Reference

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

EDR (1993a). EDR Electronic Dictionary Specification Guide.*TR-041*.

EDR (1993b). English Word Concept Dictionary. *TR-026*

Komatsu, E., Cui, J. and Yasuhara, H. (1993). A Mono-lingual Corpus-Based Machine Translation of the Interlingua Method. *Fifth International Conference on Theoretical and Methodological Issues*

Nagao, M. (1984). 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.*

Sadler, V. (1989). Working with Analogical Semantics, *Disambiguation Techniques in DLT*, Foris Publications, Dordrecht Holland.

Sato, S. (1991). 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.

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