Mechanical Translation from Japanese into English

TOSHIYUKI SAKAI*, SHIGEHARU SUGITA* AND AKIMASA WATANABE*

Abstract

This report describes a method and experimental results of mechanical translation from Japanese into English by a medium size digital computer. Input sentences, which are fed into the computer in separate phrase form in Japanese "KATAKANA" letters, are analyzed with the "Table Look-up" method. It generates psude-English sentences which include operators such as ED- or DO-. By using transformation rules on this sentences, normal orthographic English sentences are obtained.

For experiment, we used 200 Japanese idioms, 8,000 word heads, and 600 rewriting rules. Any form of input sentences are accepted.

1. *Introduction*

The beginning of mechanical translation between natural languages by digital computer can be traced back to 1946. Since then many trials have been performed especially for Indo-European languages. As for Japanese a few groups have tackled English into Japanese translation since 1958 in Japan. We have been one of such groups, and obtained considerably successful algorithm for machine oriented English-Japanese translation. All rewriting rules are written in context-free form, so our system may be interesting from the linguistic point of view.

This paper describes the Japanese-English translation method principally based on E-J translation which was developed by the same authors. The characteristic points of this method are that (1) model of language is based on the phrase structure grammar, (2) all rewriting rules have context-free form and they are ordered according to their role in sentences, (3) a kind of transformation grammar is introduced, (4) the algorithm is very simple and processing time is short, (5) the grammar is separated from program so that the ability of this system can be easily enlarged only by the addition of new rewriting rules.

2. *Linguistic feature of Japanese*

2.1 *Structure of Sentence*

Sentence is consisted of phrases, and needs at least a predicative phrase. Each element can be defined as follows;

This paper first appeared in Japanese in Joho-Shori (the Journal of the Information Processing Society of Japan), Vol. 10, No. 6 (1969), pp. 418–427.

* Faculty of Engineering, Kyoto University.

⟨Independent word⟩ = ⟨Substantives⟩ | ⟨declinables⟩ | ⟨adverb⟩ | ⟨qualifier⟩ |
⟨connective⟩
⟨Dependent word⟩ = ⟨particle⟩ | ⟨auxiliary verb⟩ | ⟨dependent word⟩ +
⟨dependent word⟩
⟨Phrase⟩ = ⟨independent word⟩ | ⟨independent word⟩ + ⟨dependent word⟩
⟨Sentence⟩ = ⟨phrase⟩ | ⟨sentence⟩ + ⟨phrase⟩

In Japanese, there are no fixed sentence patterns, but next rules restrict in a certain sense the structure of Japanese:

(1) fundamental constituents of sentences are phrases,
(2) predicative phrases appear at the end of sentence,
(3) subject is a kind of verb modifier, and it is not an indispensable constituent for a sentence,
(4) complex sentences can be constructed freely from predicative phrases of noun modifier form,
(5) modifier usually comes before the word which is modified by it.

As far as above rules are kept, the phrase order is fairly arbitrary.

2.2 *Part of speech*

Classification of part of speech for mechanical translation must be performed by considering the role of words both in Japanese and English. Examples of part of speech for independent words, dependent words, and word groups or cluster are given in Table 1, Table 2, and Table 3.

Dependent words play a very important role in Japanese; they indicate the word order or words to be inserted in English when they are translated. They must be classified detail enough so that structure analysis can be performed correctly and easily.

Cluster symbols are given to a group of words which have special function in the sentence, and are treated as the same level with that of independent or dependent words.

2.3 *Grammar*

A grammar for mechanical translation is represented by a set of rewriting rules used in phrase structure grammar. There are three forms;

$$\alpha\ \beta\ \gamma\ \delta \rightarrow \varepsilon\ (\xi, \sigma)$$
$$\alpha\ \beta\ \gamma \rightarrow \varepsilon\ (\xi, \sigma)$$
$$\alpha\ \beta \rightarrow \varepsilon\ (\xi, \sigma)$$

where $\alpha, \beta, \gamma, \delta$ and ε are symbols for part of speech, ξ is an indicator for word order in English, and α indicates words to be inserted. The meaning of rewriting rule is that if the same sequence of part of speech on the left hand of rewriting rule appears in the sentence, the sequence of symbols is replaced by the right hand symbol ε, and according to ξ, σ an apropriate subtranslation for that sequence is constructed (see Fig. 1).

Table 1. Examples of independent words.

Symbol	name	example	Symbol	name	example
AA	adjective	美し，はや	N 5	noun	すき，きらい
AB	〃	美しく，はやく	N 6	〃	今日，来年
AC	〃	美しい，はやい	N 7	〃	みな，みんな
B 1	adverb	たいへん，かなり	R 1		この，わが
B 2	〃	ほぼ	IA	intransitive verb	歩か (ARUK)
B 3	〃	もちろん	IB	〃	歩き
CN	conjunctive	そして，また	IC	〃	歩く
CT	〃	とき	ID	〃	歩い
N 1	noun	山，川	TA	transitive verb	読ま (YOM)
N 2	〃	あいまい，静か	TB	〃	読み
N 3	〃	私，彼	TC	〃	読む
N 4	〃	こと，ところ	TD	〃	読ん

Table 2. Examples of dependent words.

Symbol	name	example	Symbol	name	example
X 1		から，まで	XF	subjunctive	ば
X 2	posesive	の	Z 0	adjunctive	ている
X 3	subjective	が	Z 1		たい，ます
X 4	〃	は	Z 4	predicative	だ，です
X 5	objective	に	Z 7	imperative	なさい
X 6	adverbial	へ	Z 8	negative	な
X 7	〃	で	ZA	comparative	より
X 8	objective	を	ZV		のようだ
X 9	connective	と			
XA	〃	ので	. .	period	.
XB	〃	のに	# #	head, tail	
XC	〃	ても	, ,	comma	,

Table 3. Examples of cluster symbols.

Symbol	function	example	Symbol	function	example
BP	adverbial	N 1+X 1	SY	clause	MW+VY
ME	〃	N 1+N 6	SZ	〃	N 1+X 2+VY
MG	subject	N 1+X 3	UY	predicative	ID+ZY
MI	object	N 3+X 5	VA	〃	TB+Z 1
MO	〃	N 3+X 8	VP	〃	MO+VA
MW	subject	N 1+X 4	VX	imperative	TA+Z 7
RY	celaus	MW+UY	VY	predicative	TB+ZY
SG	subjective	SY+X 3	VZ	〃	TD+X 7
SO	objective	SY+X 8	VF	subjunctive	TB+XF
SS	sentence	MW+VP	KB	interrogative	KA+##
SX	imperative	##+VX			

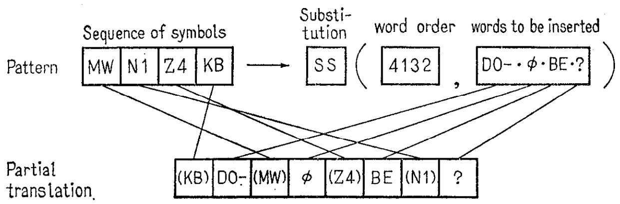

Fig. 1. Structure of pattern and partial translation.

Table 4. Examples of patterns.

class	sequence of symbols	substitution	word order	auxiliary word
A 4	## MO TB XF	MW	1 3 2 0	3 0 0 0
	N 1 ZA AA Z 8	VY	4 3 2 1	1 0 0 0
	BP CN N 1 X 5	BP	1 2 3 0	0 0 0 0
	BP VP ,, VP	VP	2 1 4 0	0 G 0 0
A 3	B 1 B 1 X 9	BP	1 2 0	0 0 0
	MG TY N 1	N 1	3 1 2	2 0 0
	MG AY N 4	N 1	3 1 2	2 1 0
	MG TY TY	SY	1 2 0	0 0 0
A 2	AB N 2	N 2	1 2	G 0
	AB VF	VF	2 1	0 0
	KA ##	KB	0 0	0 0
	B 1 AY	AY	1 2	0 0
B 4	## MG VP KB	S S	4 2 3 0	> 0 A 0
	## MO TB ZS	S S	1 4 3 2	I 0 0 A
	MW B 1 MG VA	S S	1 4 3 2	0 0 0 0
	VP MW VA ##	S S	4 1 2 3	I 0 0 0
B 3	MI MG VA	S S	2 3 1	0 0 0
	MG VP KB	S S	3 1 2	> 0 A
	MW CN VP	S S	2 1 3	0 0 0
	MW N 3 Z 1	S S	1 3 2	1 0 0
B 2	CN VA	S S	1 2	I 0
	MW VA	S S	1 2	0 0
	SS KB	S S	2 1	> A
	SS SS	S S	1 2	0 0

auxiliary words 1: BE, 2: THAT, 3: TO, A: ?, G: AND, I: I (WE, IT), >: DO-

Patterns or rewriting rules are classified into two major classes A and B according to their role in sentence. Patterns in each class are further subdivided by the length of their left hand, i. e. number of symbols on the left hand. Therefore there are six classes in total, i. e. A4, A3, A2, B4, B3, and B2. A4, for example, is a class of patterns which belong to the class A and their number of symbols on left hand is four. Some examples are shown in Table 4.

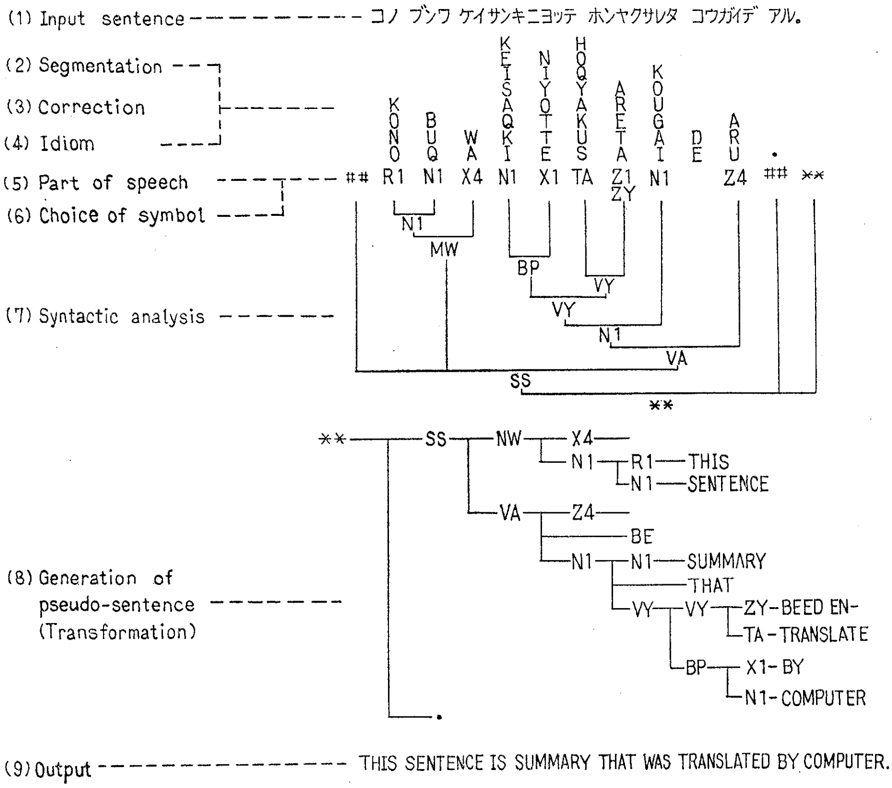

Fig. 2. Process of mechanical translation.

3. *Algorithm of translation*

Translation process printed by the computer is shown in Fig. 2.

(1) Input; Input sentence is written in KATAKANA and in the form of separate phrase.

(2) Segmentation; A typical structure of phrase is 'independent word' + 'dependent word'. So input sentence in phrase form must be segmented into word unit. In this case dependent word or words are considered as if they are suffix in Indo-European language. Suffix table consists of about 200 dependent words.

(3) Correction of segmentation; For some words segmentation always yields wrong results. This wrong segmentation is amended by the use of word table containing about 200 words which are usually segmented wrongly.

(4) Idiom; a set of words, which convey a definite meaning as a whole, must be treated as if they are single word. The idiom dictionary has about 200 sets of words.

An example of above four processes is illustrated below;

(1) KONO HOQWA ZYUUYOUNA BUQKENDE ARU

(2) KO-NO HOQ-WA ZYUU-YOUNA BUQKEN-DE ARU

(3) KONO HOQ-WA ZYUUYOU-NA BUQKEN-DE ARU

(4) KONO HOQ-WA ZYUUYOU-NA BUQKEN DEARU

(5) Word dictionary; each word is consulted with word dictionary which has about 8,000 heads, and is given parts of speech and corresponding English words.

(6) Selection of part of speech; when a word is given two parts of speech by the dictionary, one of them which is suitable for that sentence must be selected. "WATASHI", for example, has two symbols TB (as a verb) and N3 (as a pronoun). If a word having a symbol X4 followed this word, then N3 is chosen, but if a word with Z1 symbol comes after this word, then TB is chosen. Otherwise the order specified in the dictionary is adopted.

(7) Syntax analysis; according to the priority mode among the patterns or rewriting rules, simple pattern matching technique is applied to the sequence of part of speech in sentence in the way schematically shown in Fig. 3. First,

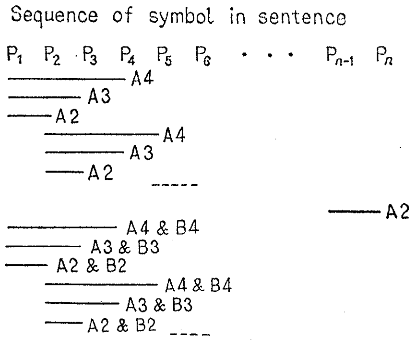

Fig. 3. Algorithm of parsing.

from the top of the sequence four symbols $P_1P_2P_3P_4$ are taken out, and this sequence of four symbols is compared with that of left hand of patterns in A4. If there is no coincident one, $P_1P_2P_3$ is compared with that in A3. If not, P_1P_2 is looked for in A2. If this matching is not successful, then $P_2P_3P_4P_5$ is taken out, and this is looked for in A4, and then $P_2P_3P_4$ in A3, P_2P_3 in A2. If $P_iP_{i+1}P_{i+2}$ is found in A3, this sequence of symbols in sentence is replaced with the right hand symbol of the coincident pattern, and corresponding partial translation is constructed (see Fig. 1). After this substitution, the same matching procedure is repeated. If there still remain several symbols in the sentence after process comes to the end of sentence, then process returns to the top of the sentence and the same procedure is continued using this time both class A and B.

(8) Generation of translated sentence; At the time when syntactic analysis is finished, a pseudo-sentence which includes operators is obtained. A pseudo-sentence is shown in Fig. 2 at the end of branches of lower tree. After some transformations are applied, the final translation sentence is obtained.

4. *Experiments*

Experiments were performed using a medium size general purpose computer; NEAC 2200/200, core 16k characters, cycle time 2μs. As a backing store a magnetic tape was used. Program was written in assembler language and had about 1,800 statements. Translation time was about 40 sec for sentence consisting of about 30 phrases. One-third of time is spent to transfer data from magnetic tape into main memory. Several examples of translated results are shown in Table 5. Distinction between single and plural form is not considered, and articles are neglected because in Japanese such information is not expressed explicitly. An expression such as I (WE, IT) is inserted when original Japanese sentence has no subject part.

5. *Conclusion*

This paper describes a trial of mechanical translation from Japanese into

Table 5. Examples of mechanical translation.

○右へまがると，郵便局が見つかるでしょう.
TURNING TO RIGHT, POST OFFICE WILL BE FOUND.

○国境の長いトンネルを抜けると雪国であった.
PASSING THROUGH LONG TUNNEL OF BORDER, IT WAS SNOW COUNTRY.

○あなたはすぐ駅へ行ったほうがよい.
YOU HAD BETTER GO TO STATION AT ONCE.

○私があなただったら私はそれをしただろう.
I WAS YOU I WOULD DO IT.

○私は馬鹿だ，しかし，私は愛することを知っている.
I AM FOOL BUT I KNOW TO LOVE.

○私は少年だ．そしてあなたは少女です.
I AM BOY AND YOU ARE GIRL.

○彼女は父によって与えられたリンゴを食べている.
SHE EAT APPLE THAT WAS GIVEN BY FATHER.

○我々は彼が昨日会った男を知っている.
WE KNOW MAN THAT HE SAW YESTERDAY.

○屋根が赤い家は学校です.
HOUSE WHOSE ROOF IS RED IS SCHOOL.

○我々は水が流れている川を渡る.
WE CROSS RIVER THAT WATER FLOW.

○私は彼が正直であることを知っている.
I KNOW THAT HE IS HONEST.

○音が空気の振動であることは読者諸兄のよく知るところと思うが,しかし,我々が話しをしたり,音楽を楽しんだりする音は,我々が使用する空気振動のごく一部分にしかすぎない.
I (WE, IT) THINK THAT READER KNOW WELL THAT SOUND IS VIBRATION OF AIR, BUT SOUND THAT WE DO STORY AND ENJOY MUSIC IS ONLY ONE PART OF AIR VIBRATION THAT WE USE.

○ここで，これからお話しする超音波は，もはや，我々の耳には聞えない周波数の高い振動である.
SUPER SONIC TO TELL HERE FROM THIS IS HIGH VIBRATION OF FREQUENCY THAT IS NOT HEARD, NOW, IN (AT) OUR EAR.

○ON-LINE TIME SHARING SYSTEM では人間と機械が直接会話することができるが，その文法は実際にはまだ不完全である.
HUMAN BEING AND COMPUTER CAN CONVERSE WITH DIRECTRY IN ON-LINE TIME SHARING SYSTEM, BUT THE GRAMMAR IS INPERFECT INDEED YET.

○自然言語と計算機言語との間のギャップをうめるためには人工の人間言語を創造することが望ましい.
TO CREATE ARTIFICIAL HUMAN BEING LANGUAGE IN ORDER TO BURY GAP OF THIS NATURAL LANGUAGE AND COMPUTER LANGUAGE IS DESIRABLE.

English. The analysis of this results tells us that a model of phrase structure grammar is applicable even to Japanese which is usually thought to be very different from Indo-European languages. Also it is useful to mention that rewriting rule of context-free form having priority mode is very effective for mechanical translation, though language itself is context sensitive. Our technique developed for analysis of sentences is not only applicable for mechanical translation, but it may be effectively used for any situation where on-line conversational communication system using natural language is adopted.

As for the next step, it is very important to investigate how to introduce semantical information into syntactic one.

References

[1] T. Sakai, and S. Sugita, Mechanical translation of English into Japanese, *J of IECEJ*, *49*, 2, 46-53 (1966).

[2] S. Sugita, A study on mechanical translation from English into Japanese, Doctoral Thesis, Kyoto University, Jan. 1968.

[3] T., Sakai, S. Sugita and A. Watanabe, Mechanical translation of Japanese into English, *Report for Information Theory Group of IECEJ*, Jan. (1969).