[*Information processing: proceedings of the International Conference on Information Processing*, Unesco, Paris 1959] 163

English-Japanese machine translation

By S. Takahashi, H. Wada, R. Tadenuma and S. Watanabe, Electrotechnical Laboratory, Tokyo (Japan)

A fully transistorized computer for the purpose of experimental machine translation from English to Japanese was recently completed, and a program test is now being conducted. This paper describes the organization of this special purpose computer named Yamato as well as the translation principles and the program flow diagrams.

Yamato, whose switching elements use 600 transistors and 7,000 germanium diodes, can work with both variable and fixed word length, executes 46 different instructions of a one-plus-one address code and stores data on a magnetic drum holding 820,000 bits. This drum holds four dictionaries: word, idiom, syntax, and Japanese word, as well as a translation program of about 4,000 words of 32 bits. Upon consulting the word dictionary, each word of a given sentence is transformed into an eight character item of information which includes its grammatical features and the location of the corresponding Japanese word. Since the grammatical structures of the two languages are quite different, word for word translation, such as might be possible between European languages, is generally unsuccessful. The program described is designed to find the grammatical structure of the given English sentence and to transpose the words in it so as to meet the corresponding Japanese grammar. After the transposition is made, each eight character word is translated into Japanese with the use of the fourth dictionary, and then typed out.

It takes about ten seconds to translate a simple sentence such as "I have some eggs in my hand", including the time required for input and output.

1. Introduction

Machine translation has already been tried at several institutions and in most cases general purpose electronic computers with ample storage capacity have been used for this purpose. The only exception may be the machine of the University of Washington [1]. In Japan, the necessity of machine translation is probably more intense than in other countries, because Japanese people have particular difficulties in learning foreign languages, due to their quite different letters and to the unique grammatical structure of their language.

The usual difficulties of machine translation are also found in the programming for translation from English to Japanese. It would be better, therefore, to examine translation principles thoroughly, using a general purpose computer, before constructing a special purpose machine. In Japan, however, computer development is still in an early stage, and until recently only a few computers, all with a storage capacity less than 1,000 words, were available. Therefore, it was decided to construct a special purpose machine, which has a relatively large magnetic drum store and handles words of variable length, but which has neither multiplication nor division mechanisms. This machine was completed about six months ago and named "Yamato," which meant "Japan" in ancient times.

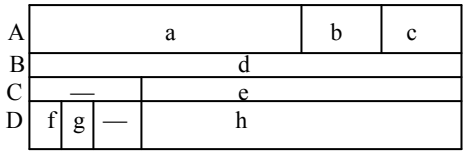
At the same time, the textbooks on English which were being used in the first and second grade classes of some Japanese junior high schools were investigated. 2,000 English words were picked out, and a program flow chart to translate the whole textbook of the first year grade was prepared. At this stage neither relative pronouns nor relative adverbs appear and present perfect tense has not yet been used, but this would be the pertinent stage for the first trial of English-Japanese machine translation. A test of the program on Yamato has now been conducted. This paper describes translation principles and program flow diagrams at this stage as well as the organization of Yamato.

2. Dictionaries and tables

Four kinds of dictionary: word, idiom, syntax, and Japanese word dictionaries are stored on the magnetic drum. In the dictionaries the length of each separate item, an idiom for instance, is naturally variable. Three kinds of table, in which each item is a fixed-length code-word of eight characters each of eight bits, are also stored on the drum. These correspond to the word, idiom and syntax dictionaries, and are called the word, idiom and syntax tables respectively.

2.1 The word dictionary is simply an arrangement of 2,000 English words in the order of probable frequency of use. The word table, on the other hand, is an arrangement of eight character code-words which correspond to the contents of the word dictionary, one by one and in the same order. Each code-word indicates the grammatical features of the corresponding English and Japanese words, the location of the latter in the Japanese word dictionary, etc. By consulting the word dictionary, each word of the given English text is transformed into the address of the corresponding eight characters in the word table.

Fig. 1 indicates the structure of such an eight character code-word. In the figure A, B, C and D each consist of two characters of eight bits, one bit of which is a parity check bit and has been omitted for simplicity of explanation. In the eight character code-word there are a number of separate items indicated by a, b, etc., a, d and e indicate the English part of speech, the Japanese part of speech and the location of the corresponding Japanese word, respectively, b and c



The first meaning of "spring"

A	0 0 0 0 1 1 0 0 0 0 0 0 0 0
B	1 0 0 0 1 1 0 0 0 0 0 0 0 0
C	0 0 0 0 1 0 1 1 1 0 1 1 1 1
D	0 1 0 1 0 1 0 0 1 0 1 1 0 0

The second meaning of "spring"

A	0 0 1 1 1 1 0 0 0 0 0 0 0 0
B	0 0 0 0 0 0 0 0 0 0 0 0 0 0
C	0 0 0 0 1 0 0 1 1 0 1 0 1 0
D	0 0 0 0 0 0 0 0 0 0 0 0 0 0

Fig. 1. Contents of eight character code-word with an example. A, B, C and D are two characters each.

are the locations reserved for the information concerning the affixes which are removed from the original word when the word dictionary is consulted, c indicates the existence of an affix and b its type. The bit f denotes whether the word occurs with other items in the word dictionary in at least one idiom. The bit g denotes whether the word has another meaning of different part of speech or not and h gives the location of another eight character word corresponding this meaning in the word table.

In fig. 1 the word "spring" is shown as an example. Two meanings of "spring" appeared in the textbooks investigated; one is the intransitive verb which means トビハネル (tobihaneru), that is "leap" or "jump," and the other is the noun which means ハル (haru), "the season of the year." Fortunately, the noun which means バネ (bane), "un ressort" in French, has not appeared. At the present stage multiple meaning for the same part of speech must be avoided since the program storage capacity is limited.

For the first meaning in the example, the two characters represented as A indicate that it is a root of a perfect intransitive verb; B, that its Japanese equivalent is a verb which is conjugated following a rule indicated; C, that its Japanese equivalent is the 751st word in the Japanese dictionary; D, that it does not occur in any idiom, but that it has at least one other meaning, which is given in the 1324th location of the word table.

For the second meaning, A denotes that it is a common noun; B and C, that its Japanese equivalent is a common noun and the 618th word in the Japanese dictionary: D, that it has no further meaning.

2.2 The idiom dictionary is an arrangement of groups of words, also in the order of frequency of use. Each word is represented by the two characters D denoting the location of any other meaning. For the word which has no other meaning but occurs in one or more idioms, a number denoting a pseudo-location is given. Whenever more than two words having "1" in the bit f appear successively, this dictionary is consulted. The biggest group of words which coincides with a content of the dictionary is assumed to be an idiom in the given sentence, and is changed into the address for the corresponding eight character word in the idiom table. Idioms which consist of words separated from each other, such as "so ... that ..." are excluded from this dictionary and must be treated by a program. - %

2.3 The syntax dictionary is an arrangement of *20* groups of parts of speech, corresponding to the part A of the code-words. The syntax table consists of addresses indicating the beginning of the program subroutine that should be used for the corresponding syntax.

2.4 The Japanese word dictionary is simply an arrangement of Japanese words.

3. Translation principles and flow diagram

Fig. 2 shows in outline the principles of English-Japanese machine translation. A given sentence is read in word by word, and changed into a series of eight character code-words with the use of the word dictionary and table. Any word which cannot be found in the word dictionary, even after the removal of affixes, is put in a separate track of the magnetic drum for the purpose of direct type-out (transliteration). It is assumed to be a noun, and is also replaced by an eight character code-word. The idiom dictionary is used after a whole sentence has been read in.

Since the grammatical structures of the English and Japanese languages are quite different, "word for word" translation, such as might be useful for translating between two European languages, is generally unsuccessful. It is important to find the grammatical structure of a given English sentence first, and then to transpose the words according to the corresponding Japanese grammar. The syntax dictionary is consulted for this purpose.

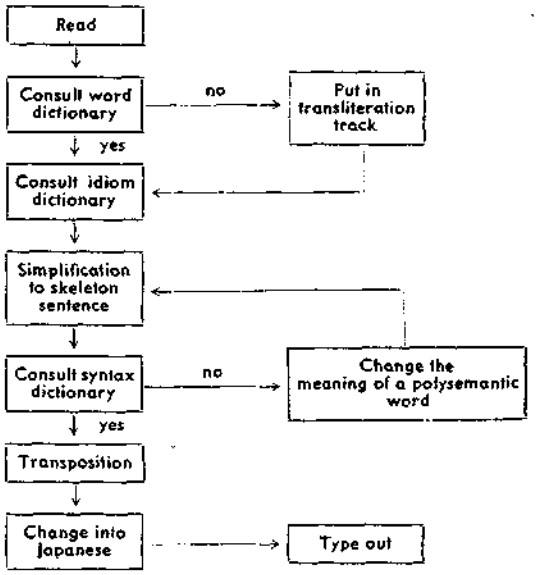


Fig. 2. Translation principle

Before consulting the syntax dictionary, however, it is necessary to simplify the given sentence to the form of a basic pattern which is included in the dictionary; "noun + transitive verb + noun", for instance. The procedure for doing this will be described later. If the simplified pattern for the given sentence is found in the syntax dictionary,

then the subroutine for the word transposition is given by the syntax table, as stated in section 2.3. If it is not found, the meanings of the polysemantic words are changed one at a time, and the simplification processes are repeated from the beginning. The multiple meanings are arranged in a sequence which considers each word first as an auxiliary verb and, finally as a noun, in order to prevent endless iterations. If the pattern cannot be found by any means, the sentence is translated word for word.

The subroutines not only transpose the words, but also insert certain words which are peculiar to the Japanese language. After this, every group of words is decomposed into the original eight character code-words, and each word is changed into Japanese with the use of the Japanese word dictionary. Finally the words are typed out one by one. Japanese sentences are generally written in a mixture of three kinds of letter; 'kanji' (Chinese ideographs), 'hirakana' and 'katakana'. In the present experiment only 'katakana', consisting of 75 letters is used. The alphabet (capital letters only) also can be typed out, and is used for transliteration.

A very simple example of the translation processes is shown in fig. 3, and the program flow diagram in fig. 4. It takes about ten seconds to complete the translation of fig. 3, including the time required for the input and output. The whole translation program of fig. 4 requires about 4,000 instructions.

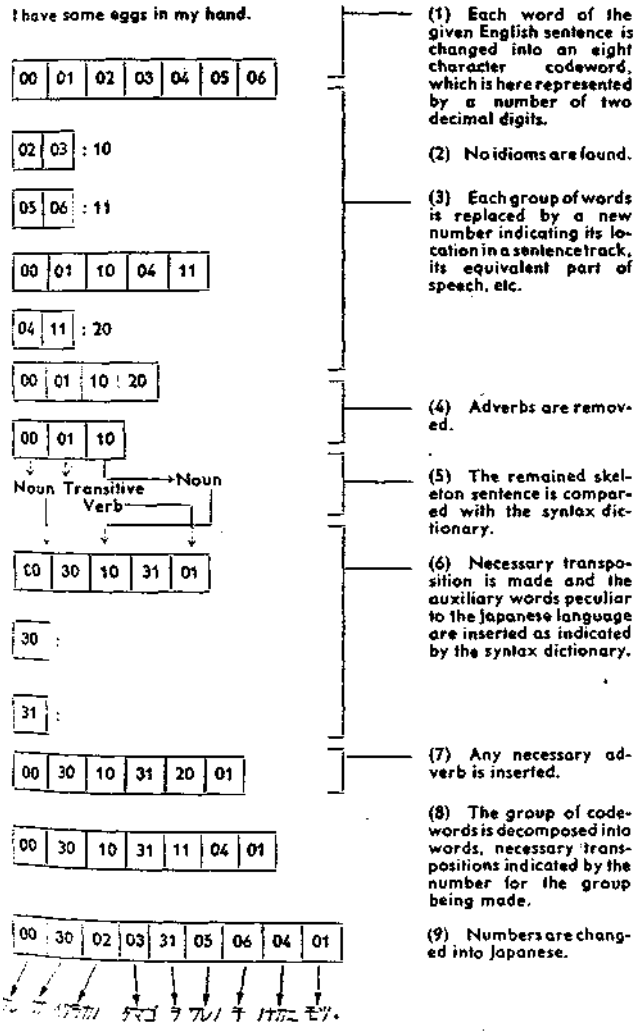


Fig. 3. A simple example of English-Japanese Machine

The process of simplification to the basic pattern is shown in fig. 4. The process may be divided into three parts; 1) grouping of words, 2) decreasing the number of verbs to one, 3) removing adverbs. The last part is simple and requires no explanation. The second part is described fairly precisely in fig. 4. The first part is subdivided into five types of grouping; (a) auxiliary verb + verb → verb, (b) adverb + adjective → adjective, (c) adjective or possesive form + noun → noun, (d) article + noun → noun, and (e) preposition + noun → adverb. In the case of the preposition "of", the group (e) often reduces to an adjective modifying the preceding noun. Therefore, "of" is treated separately and the preceding word is checked.

4. Machine organization

Yamato is a binary serial computer and operates at a clock repetition rate of 195 kc. Instructions and numerical word are both represented by 32 bits, including one bit for a parity check. A magnetic drum of 200 tracks, having a capacity of 820,000 bits and an average access time of 10 milliseconds, is used as the store. A one-plus-one address code is employed to compensate for the slow speed of the drum. The address parts of the instruction are 12 bits each, and 6 bits are used for the functional part.

Yamato can execute 46 different instructions, detailed in Table 1. Each instruction has two address parts, A_1 and A_2. The location of the next instruction is always indicated in A_2, except in the case of jump instruction. For some instructions A_1 indicates an address in the program section of the store while for others it supplements the functional part in the designation of the operation. In the latter case, A_1 is subdivided into three hexadecimal numbers A_{11}, A_{12}, A_{13}

Since Yamato is a special purpose computer, some of the instructions are peculiar to it. It handles information of variable length, such as an English word which may have up to 16 characters for which two letter registers of 8 characters each are used for this purpose. There are also four counters to count word or letter numbers automatically in some operations.

The storage is divided into three large sections; dictionary, table and program section. The dictionary section, which includes the tracks for the words to be transliterated, stores information of variable length. Neighboring words are separated from each other by an "all mark," namely a character whose eight bits are all 1. When a word is stored in this section, the "all mark" is automatically inserted directly behind the word. The address of a word in this section is determined by counting the "all" marks from the top. The table section, including the tracks which store the sentence in various forms as it is treated, stores only eight character code-words. The program section has 4096 locations and stores instructions and numbers of 32 bits. The English text is presented in the form of punched tape. This is punched by hand at present, but in the near future the punching will be performed automatically by a reading machine [2].

Being completely transistorized, Yamato consumes only 50 watts excluding the power for the drum motor and for the input and output. It employs dynamic basic circuitry, including a one bit delay which was invented by Takahashi, one of the authors, and which has been successfully operating in a general purpose computer, ETL Mark IV [3] for more than a year.

5. Conclusion

As the program test is not finished, it is difficult to draw any concrete conclusions on the English-Japanese machine translation. However, the experience of programming shows that there are no essentially difficult problems, ex-

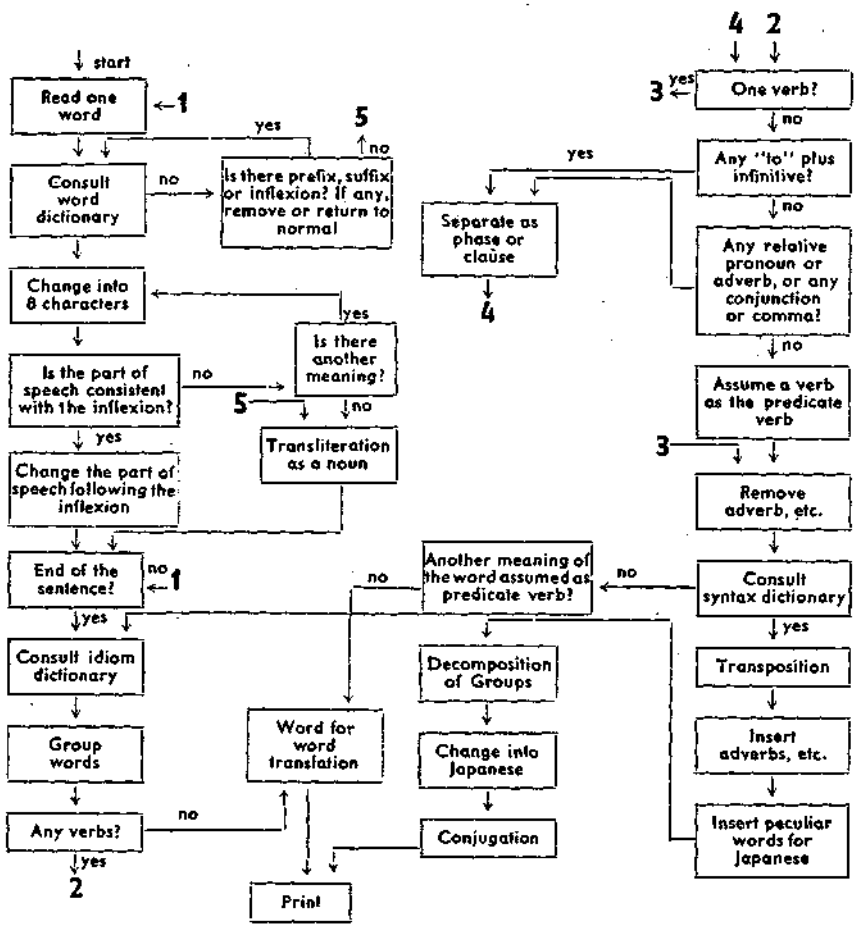


Fig. 4. Flow diagram

cept when a word has multiple meanings in the same part of speech. In comparison with the translation between European languages, two problems predominate; the word order difficulties and the problem of auxiliary words peculiar to: Japanese, which have no corresponding part of speech in English.

As for the computer Yamato, it will be necessary to increase the storage capacity in the near future. It is being planned to add a photographic permanent store such as is used in the machine of the University of Washington [1] for dictionaries and tables.

The Japanese language, which has been cultivated in an island-country for many years, was largely modified by the introduction of Chinese letters about ten centuries ago. It is felt that there are so many irregularities that the language, and also the letters, have to be modified to ease machine translation. A similar request may arise also for English. It is desirable for the mutual understanding of all the peoples on the globe that articles in various languages should be written according to rules which are convenient for machine translation.

6. References

[1] WALL, R. E.: University of Washington Eng. Exp. Station Rep., No. 108, 1956.

[2] WADA, H. et al: *An electronic reading machine,* these proceedings *IV* B.

[3] NISHINO, H. et al: *ETL Mark IV, a transistor digital automatic computer* Journal of the Institute of Electrical Communication Engineers of Japan 1959, to be published.

Table 1. Instructions of Yamato

Code	Abbreviation	Operation
02	Add	The number in A_1 is added to the accumulator
03	Clear Add	The number in A_1 replaces the contents of the accumulator
04	Sub	The number in A_1 is subtracted from the accumulator
05	Clear Sub	The negative of the number in A_2 Replaces the contents of the accumulator
06	Store	The contents of the accumulator are copied to A_1
07	Clear Store	Both the accumulator and A_1 are cleared
10	Read in to Acc	One character on the tape is read to the accumulator. The previous contents of the accumulator are shifted A_{13} places to the left
11	Clear Read in to Acc	The accumulator is cleared. The other actions are the same as in 10
12	Read in to LR	One English word, not longer than 16 characters, is read to the letter register
14	Acc Shift	The contents of the accumulator are shifted A_{13} places; to the left when A_{11} is even, and to the right when A_{11} is odd
20	Raise Acc	A_1 is added to accumulator
21	Clear Raise Acc	A_1 replaces the contents of the accumulator
22	Lower Acc	A_1 is subtracted from the accumulator
23	Clear Lower Acc	—A_1 replaces the contents of the accumulator
24	Add Counter	The contents of the counter designated by A_{13} are added to the accumulator
25	Clear Add Counter	The contents of the counter replaces the contents of the accumulator.
26	Acc to Counter	A part of the accumulator contents replaces the contents of the counter designated by A_{13}
27	Clear Acc to Counter	Both the counter designated by A_{13} and the accumulator are cleared

Table 1 (continued)

Code	Abbreviation	Operation
30	Add Letter	The A_{11}th letter in the letter register is added to the accumulator
31	Clear Add Letter	The A_{11}th letter replaces the contents of the accumulator
32	Acc to LR	The 7 least significant bits of the accumulator replace the A_{11}th letter of the letter register
33	Clear Acc to LR	Both the accumulator and the A_{11}th location of the letter register are all cleared
34	Extract to Acc	The logical product of the 7 least significant bits of A_1 and the A_{11}th letter of the letter register is brought to the accumulator. The previous content of the accumulator is shifted 7 places to the left
35	Clear Extract to Acc	The accumulator is cleared before the logical product stated above is entered
40	Bring from Table to LR 1	An eight character code-word is brought to the letter register 1 from the location in the table section of the store which is defined by A_{11}, A_{12} and the contents of the counter designated by A_{13}
41	Bring from Table to LR 2	An eight character code-word is brought to the letter register 2 from the location defined in 40
42	Store LR 1 to Table	The contents of the letter register 1 are copied to the location defined in 40
43	Store LR 2 to Table	The contents of the letter register 2 are copied to the location defined in 40
50	Bring from Dictionary	A word is brought to the letter register from the location of the dictionary section of the store which is defined *by* A_{11}, A_{12} and the contents of the counter designated by A_{13}
51	Store to Dictionary	A word in the letter register is stored to the location defined in 50

Code	Abbreviation	Operation
52	Consult WD	The contents of the letter register are compared with the contents of the word dictionarv. If coincidence is obtained, next instruction is taken from A_2 and the address of the word in the dictionary is left in a particular counter called the dictionary counter. If not, next instruction is taken from A_1
53	Consult ID	The comparison is with the idiom dictionary. The other actions are the same as in 52
54	Consult SD	The comparison is with the syntax dictionary. The other actions are the same as in 52
55	Consult X	Not used
56	Consult Y	Not used
60	Raise Counter	The content of the counter designated by A_{13} is increased by one
61	Lower Counter	The same counter is reduced by one
62	LR Shift	The contents of the letter register are shifted by A_{13} characters; to the left when A_{11} is even, and to the right when A_{11} is odd
63	LR 2 to LR 1	The two letters in the letter register 2 designated by A_{11} are brought to the most significant digits of the letter register 2
64	Acc Minus Jump	The next instruction is taken from A_1 if the content of the accumulator is negative. Otherwise, from A_2
65	Acc Zero Jump	The next instruction is taken from A_1 if the content of the accumulator is zero. Otherwise, from A_2
70	Type out LR	Type the contents of the letter register
71	Type Special Char.	Type the character designated by A_1
72	Clear Counter	The counter designated by A_{13} is cleared
73	Stop	The machine is stopped