Stenograph to Japanese Translation System

Hiroshi Kinukawa,* Kenji Tsutsui,* Ikuo Odagiri** and Mutsuko Kimura***

Abstract

A stenograph to Japanese (Kanji-Kana representation) translation system, which is one of the Japanese Information Processing systems, is introduced in this paper. Input data to this system are phonetic symbols expressing spoken Japanese and are not divided into Japanese words. The main feature of this system is that it translates the above-stated input string into normal Japanese sentences, and outputs them in combination of Chinese character, i.e. Kanji, and Kana alphabet. We have adopted a method of morphological analysis as translating algorithm and have made the dictionary by investigating spoken Japanese words. The mistranslating-ratio was 4% and the translating ratio of Chinese borrowed phonetic words into Kanji was 67%. The processing time was 155 msec/word.

1. Introduction

The increasing number of people, needed in preparation for input or output of data to be processed by computer, have brought us the demand of human-oriented method of Japanese language manipulation system. Usually written Japanese is presented by the combination of Chinese character, i.e. Kanji, and phonetic Kana. A couple of devices have been developed for outletting the Kanji data from computers, but inputting devices, or input methods of Kanji data are under-developed because of the difficulty of handling not a few kinds of Chinese characters we use. We daily use more than 2,000 Chinese characters. This report is concerned with one of the input methods of Japanese language processing, and we introduce the system of translating stenograph into Japanese, which we developed since 1971. In this stenography, we use stenotypes.

2. Characteristics of stenograph and Japanese output

2.1 Characteristics of stenograph by a stenotype

The stenographic symbol string is one kind of sound-string which is typed continuously by a stenotype. This stenograph has typing rules as follows for efficiency. These are different points from normal Japanese.

(1) Abbreviated expressions of idioms.

(2) Existence of symbols which have more than two sound strings.

ex) A symbol expresses a sound string [jo] and a sound string [gozonji].

(3) Expressions of voiced sound words by using clear sounds.

ex) 寝言 [negoto] ——→ [nekoto]

This paper first appeared in Japanese in Joho－Shori (Journal of the Information Processing Society of Japan), Vol. 16, No. 6 (1975), pp. 484～491.

* Systems Development Lab. Hitachi, Ltd.
** Facom Hitac Ltd.
*** Institute of Behavioral Sciences

(4) Existence of symbols which have sounds in parentheses as a part of expressing sounds.

ex)((シ)テ) クル 〔((shi)te)kuru〕;

This expresses 〔shitekuru〕, 〔tekuru〕 or 〔kuru〕 by context.

(5) A stenographer can express a verb or an adjective connected with an inflected function word by lining up symbols which are original forms of both words.

2.2 Japanese output

This system must process following points by reason of characteristics of Japanese.

(1) Translation of Chinese-borrowed phonetic words to Chinese characters.

(2) Translation of foreign words and words of foreign origin to "Kata-Kana".

(3) Translation of sounds which express numbers to Chinese characters.

(4) Translation of function words 〈o.e.wa〉 to <を・へ・は>

(5) Automatic Insertion of punctuation marks.

(6) Generation of normal Japanese sentences by processing typing-rules which were given in 2.1.

2.3 Translation problems

The system has two difficult points to process as follows.

(1) Segmentation of strings of phonetic symbols of a run-on style to unit-words.

(2) Discrimination of homonyms not only in the Chinese character translation but also according to the characteristic of the stenography.

3. Translation Method

3.1 Recognition of unit-word

It is considered that a continuous phonetic symbol string is partitioned to words by using longest-matching method. But generally several segmentations of phonetic-symbol string to unit-words exist and homonyms sometimes exist. Then it is necessary to analyze connective relation of adjacent words for correct segmentations. We adopted morphology as informations of analyzing connective relations. This method is called morphological analysis.

3.1.1 Classification of words

We classified words according to Japanese grammar, typing rules of a stenotype and our studies on the vocabulary of conversation. They are classified to 203 classes.

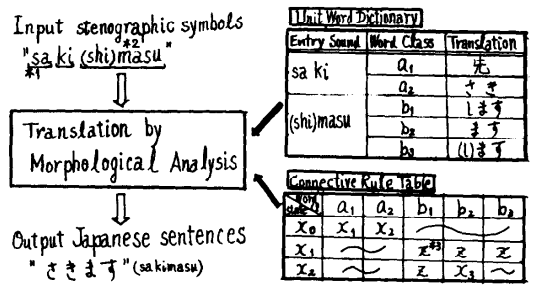


Fig.1 Algorithm of Morphological Analysis

3.1.2 Connective rule table

Connective rules of word classes are formalized into a finite state grammar and are arranged into a transition table. Terminal symbols are word classes in 3.1.1. Non-terminal symbols, that is states, are grammatical states of phrases which are

composed of several words. Number of states is 99. They contain states which mean insertion of punctuation marks.

3.1.3 Unit word dictionary

The unit word dictionary is a machine dictionary which is used for recognition of words from a stenographic symbol string.

3.1.4 Algorithm of morphological analysis

We will explain the algorithm of morphological analysis by using Fig. 1. Input is assumed for example to be a stenographic symbol string "sa.ki.(shi) masu". Both stenographic symbol strings "saki" and "(shi) masu" have homonyms in the unit word dictionary. Word classes are expressed in a_i and b_j. The smaller subscript i and j is, the higher the priority of the word is. x_k means a state, x_0 is the initial state and z is the impossible state of connection.

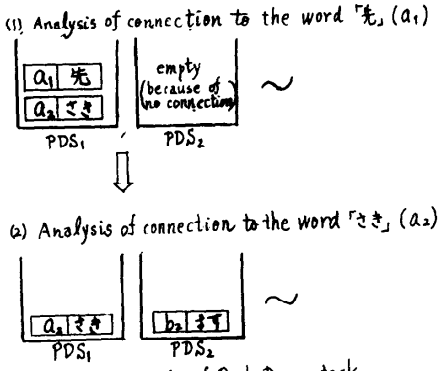


Fig.2 An Example of Push Down stack

(1) "a_1" is gotten for "saki" by searching the dictionary. Analyses for "a_1" are $x_0 \longrightarrow a_1x_1$, $x_1 \longrightarrow b_1z$, $x_1 \longrightarrow b_2z$, $x_1 \longrightarrow b_5z$. Then none connects.

(2) Secondly, analyses for a_2 are $x_0 \longrightarrow a_2x_2$, $x_2 \longrightarrow b_1z$, $x_2 \longrightarrow b_2x_3$. Then " ます " which belongs to the word class b_2 is connected to "さき". Consequently " さきます " is output as a result of translations. The result is a well-formed Japanese sentence. The process is done continuously and efficiently by introducing ten push-down stacks. The examples are shown in Fig. 2.

3.2 Translation of numbers to Chinese character

In this stenography, number data are expressed by a phonetic symbol string. These expressions need to be translated into Chinese characters. In Japanese language phonetic expressions of numerals are transformed according to vocal sounds of them when they are followed by counting units. We will explain this fact by using Table 1.

Note 1 numbers: ⟨ichi⟩ (one), ⟨ni⟩ (two), ⟨san⟩ (three)

2 counting units ⟨ban⟩ (It expresses the order.)
⟨fun⟩ (It expresses the minute.)
⟨hon⟩ (It expresses the number of things like sticks.)

We made use of this fact in the recognition of numerals.

3.3 Transformation to normal Japanese sentences

3.3.1 Inflection procedure

Each of original forms of frequent verbs and adjectives in the conversation can be expressed in one stenographic symbol. When this expression is followed by function words, it must be inflected according to the classes of function words. We classified verbs, ad-

Table 1. Examples of numeral transformation

number \ unit	ban	fun	hon
ichi	ichiban	ippun	ippon
ni	niban	nifun	nihon
san	sanban	sanpun	sanbon

*notes
~~~ : transformed part of the number
— : transformed part of the counting unit
~~~

jectives and function words according to the type of inflection.

3.3.2 Removing procedure of parentheses

This procedure is for stenotypes which have sounds in parentheses as a part of expressing sounds. The sounds in parentheses are adopted or rejected according to the relation of the preceding words.

3.4 Correction of typing errors

Stenographers make typing errors 2~3 % of all the data. Then they must be corrected. They are corrected by the off-line device which has copying function from a caset tape to another and displaying function.

3.5 Stenographer's personal shortened form

It is necessary to process each stenographer's personal shortened forms. This is able to be done well by containing these forms to the unit word dictionary. It is considered that personal shortened forms can be input together with regular forms or personal dictionaries may be set up.

4. System Organization for experiment

4.1 Computer system organization (Fig. 3)

The experiment of this system was done by using Hitachi computer system HITAC-8400.

4.2 Software system Organization (Fig. 4)

Programs of this software system are described in the assembler.

5. Results of experiment and studies of them

We made an experiment of this system on actual conversation data for the purpose of evaluating the performance.

(1) Data : (3500 words/datum) x 4

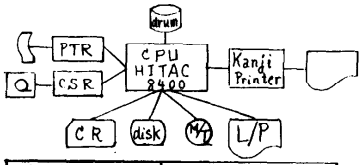
(2) The unit-word dictionary contains about 12,700 words.

(3) The result of the experiment are given in the Table 2.

Definition Exactly translating ratio

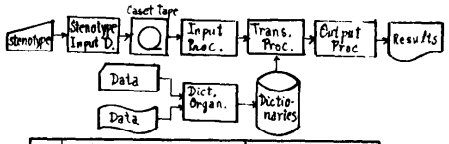
$$\overline{\overline{\text{def.}}}\ \left(1 - \frac{\text{number of mistranslated words}}{\text{total number of words}}\right) \times 100$$

$$= 100 - (\text{mistranslating ratio})$$



Device	use
Caset Tape Reader	Reading of data
Magnetic drum memory	Unit Word Dictionary
Magnetic disk memory	working
Magnetic Tape	working
Kanji Printer	Outletting of results
Card Reader	Dictionary data
Paper tape Reader	Dictionary data
Line Printer	Outletting of messages
Stenotype Input Device	Writing data to caset tape
Error Correcting D.	Correcting of errors

Fig.3 System Organization



NO.	Procedure	Number of steps (Kstep)
1	Input Procedure	3
2	Translation Procedure	7
3	Output Procedure	3
4	Dictionary Organization	9
5	Supporting Programs	20

Fig.4 Software System Organization

Character translating ratio

$$\frac{\text{number of well character translated words}}{\text{def. total number of object words translation}} \times 100$$

(a) 20 % of mistranslated words are caused by existence of homonyms and others are to be saved by a bit of improvement of the dictionary, the tables and procedure.

(b) About 80 % of not-translated words to Chinese character cannot be translated because of existence of homonyms.

(c) These points mean that man-machine interaction are needed to get higher performance than this system.

(d) The processing-time was 155 msec/word by using the Hitachi computer system HITAC-8400. The time of only translation procedure was 62 msec/word. An example of translation is given in the Fig. 5.

Table 2. Translation program efficiency

Item		Morphological Ana.		Simple Longest M.M.	
exact-translation r.		96	%	94.7	%
character translation	Chinese Charac.	67	%	58	%
	Numerals	78	%	57	%
	Proper nouns	60	%	20	%
	Kata-kana	61	%	38	%
	function w. (o, e, wa)	80	%	63	%

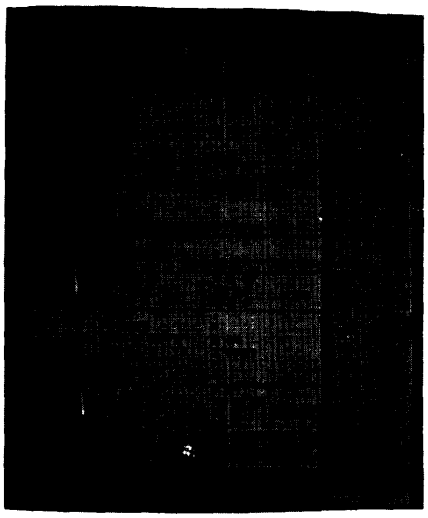


Fig. 5 An Example of Translation

6. Conclusion

This system has the following two characteristics.

(1) This system processes spoken Japanese.

(2) It translates undivided strings of phonetic symbols to "Kanji-Kana" representations.

Now, the unit-word dictionary has 12,700 words, but we need more words to be added to improve character translating ratio. The remaining research problems are as follows. The more-detailed study are needed on the vocabularies of spoken Japanese. Each stenographer often has different type of shortening typing forms. This problem may be included in this system by inputting shortened forms in the error-correcting time, or by preparing personal dictionaries. This morphological analysis may be applied to process other phonetic symbol strings.

ACKNOWLEDGMENTS The authors would like to thank Mr. Kinya Fujimoto, Mr. Yutaro Ito and Mr. Chuko Shimoda, who helped and encouraged the authors for the development of this system.

REFERENCES

1) The National Language Research Institute, Japan.
Vocabulary and Chinese characters in Ninety Magazines, Vol. 1 (1962), Vol. 2 (1963), Vol. 3 (1964).

2) The National Language Research Institute, Japan.
Studies on the Vocabulary of Modern Newspapers. Vol. 1 (1970), Vol. 2 (1971).