Invited Paper

Inputting Japanese from the Keyboard

TSUTOMU KAWADA*

The idea of operating computers in japanese was still just a dream in the 1960's. A revolutionary change came in 1978, with the development of kana-kanji conversion and the commercialization of the first Japanese word processor. In this new system, kana phonetics were input for ideographic kanjis. A word dictionary and a grammatical analysis program translated kanas into kanjis. There are only 48 Japanese phonetic symbols (kana), which makes it easy to fit them into an alphanumeric keyboard, and if they are input in romanized form, regular computer keyboards can be used as is.

I will first describe how kana-kanji conversion was developed, which will be followed by a discussion on homonym reduction that was necessary for the spread of the system. This paper will conclude with an overview of the developments in Japanese language inputting.

1. Introduction

The idea of operating computers in Japanese was still just a dream in the 1960's. One of the biggest factors standing in the way of the dream ever coming true was the number of characters used in the Japanese language, particularly the large number of kanji characters. Several different methods were developed for inputting the 6,800 ideographs used in everyday Japanese with an alphanumeric keyboard. A kanji was input one character at a time in these systems. All of these systems basically involved conversion during input into a specific code assigned to each character. Once the codes were mastered, kanji could be input at high speed, but the level of required training limited the spread of these systems to highly specialized fields.

A revolutionary change came in 1978, however, with the development of kana-kanji conversion and the commercialization of the first Japanese word processor [1], [2]. In this new system, kana phonetics were input for ideographic kanjis. A word dictionary and a grammatical analysis program translated kanas into kanjis. There are only 48 Japanese phonetic symbols (kana), which makes it easy to fit them into an alphanumeric keyboard, and if they are input in romanized form, regular computer keyboards can be used as is. Kana-kanji conversion is now the standard method used to input Japanese sentences and can be found in everything from personal computers to main frame computers.

The method has also been used for portable Japanese word processors, which have become extremely popular as home typewriters. And now, syntax analysis programs have been developed for Japanese sentences as a proofreading support tool.

I will first describe how kana-kanji conversion was developed, which will be followed by a discussion on homonym reduction that was necessary for the spread of the system. This paper will conclude with an overview of the developments in Japanese language inputting.

2. Kanji Character Input Problem

At least 6,800 different characters are needed to prepare business documents in Japanese. This requires a huge typewriter keyboard if you assign one key to each character. In fact, the keyboard of the Japanese typewriter, first developed more than 60 years ago, was about 10 times larger than that of an English typewriter. The typist had to move her hand to reach each character she types; it resulted in extremely slow input speed in comparison with an English typewriter.

To improve the efficiency of the Japanese typewriter, it was necessary to invent a touch method for typing Japanese.

A keyboard for kana typewriting consists of 48 keys. With this keyboard, a typist can type by the touch method.

Japanese typed first in kana only should be automatically converted into the expression style of modern Japanese, with a combination of kanji and kana. This method is called "kana-kanji conversion" which can realize fast typing.

2.1 Kana-kanji Conversion [3]

Japanese people conventionally write a sentence as one continuous string of characters with no space between words. Changes in character types (e.g. from kana to kanji, from kana to numeral, . . .) give useful clues to understand a given sentence. We separate a sentence into certain lengths of character strings ("bunsetsu") which are minimum meaningful units in

*A senior Manager of Toshiba R & D Center.

(a) にわにはにわにわとり ···Kana Expression
(b) にわには/にわ/にわとり ···Bunsetsu Expression
A slant line correspands to the stroke of bunsetsu key
(c) 庭には二羽鶏 ···Japanese Standard Expression
(d) garden in two hens ···word for Word Translation

Fig. 1 Various Expressions of Japanese Sentence.

Japanese sentences.

If a sentence is written only by kana characters in one continuous string, as shown in Fig. 1(a), the sentence is very difficult even for us Japanese to understand, due to the lack of separations between bunsetsus. In this case, we write a kana sentence with spaces between bunsetsus (Fig. 1(b)). A kana-kanji conversion system can translate a kana sentence of this type into a standard Japanese expression, as shown in Fig. 1(c).

A typist inputs a sentence by 48 kana keys, inserting bunsetsu keys between bunsetsus. Each input kana string is translated into proper kanji characters by a conversion program which is based on morphological analysis.

Morphological analysis needs its own word dictionary to analyze a given kana bunsetsu. The dictionary has a certain set of kana entries with their kanji strings, grammatical properties and frequencies of use of words. In morphological analysis, these dictionaries are consulted to avoid ungrammatical expressions.

All candidates for the homonyms which can not be uniquely determined in the analysis are sorted according to the order of their frequencies of use (i.e. a more popular word has a higher priority).

The typist selects one of the homonyms shown on a CRT display in the order of priority. This interactive operation can update their usage frequency information in the dictionary.

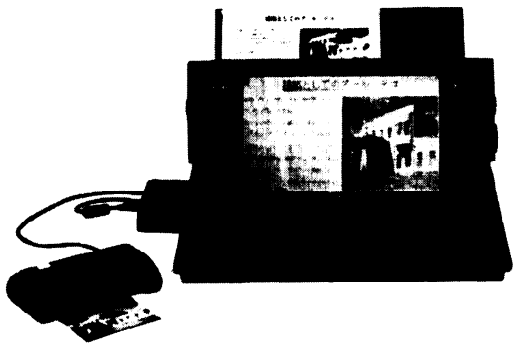
In 1978, the first Japanese word processor (Fig. 2) put the bunsetsu based kana-kanji conversion system to practical use.

The word dictionary of this word processor contained 54,000 words and 8,000 proper nouns. The main processor was a 16-bit micro-processor. It was also equipped with a dot impact kanji printer which printed a kanji character in 24×24 dots as well as line drawings.

2.2 Advances in the Kana-kanji Conversion System

In bunsetsu based kana-kanji conversion systems, it was necessary for users to press a "bunsetsu" key at the end of each bunsetsu. In new systems, however, the program determines the end of each bunsetsu in a continuous kana string, resulting in speedier, simpler keyboard input. The typist can input kana strings in a continuous form (Fig. 1(a)), compared with the bunsetsu based input string (Fig. 1(b)).

(a) the First Japanese Word Processor JW-10 (1978)



(b) Portable Word Processor (1989)

Fig. 2 Japanese Word Prosessor.

2.3.1 Kana-kanji Conversion Without the "Bunsetsu" Key

When a long kana string is to be broken up into an unknown number of bunsetsus, many combinations of conversion results are generated, with a consequently larger number of conversion candidates generated. It is necessary, therefore, to set up evaluation criteria under which candidates can be selected. The following systems have been created to do this: the "two bunsetsus longest match method [4]", the "least bunsetsu's number method [5]", and the "better method [6]".

The "two bunsetsus longest match" method was developed because the first bunsetsu had formerly been determined according to the conventional longest match method from the beginning of the kana string which sometimes resulted in mistakes cutting into the next bunsetsu. In this method, the longest candidate in a connected pair of bunsetsus (two bunsetsus) starting at the beginning of the kana string is detected, and the first bunsetsu of a connected pair is determined to be the prior bunsetsu. The kanas involved are then excluded,

and the next bunsetsu is newly searched in the same manner—by looking for a connected pair of bunsetsus in the remaining kana string. The result is a reduction in the number of bunsetsus spilling over into the next bunsetsu compared to conventional longest match methods.

The principle in the "least bunsetsu's number method" is to divide a kana string into the minimum number of bunsetsus possible. In other words, this is equivalent to looking for a combination of the longest bunsetsus for the whole kana string. In actual processing, the program looks at all possible combinations of bunsetsus for any given kana string, and chooses the combination that results in the fewest bunsetsus.

This method requires the whole kana string to start the evaluation of a combination of bunsetsus.

To improve this, the "better method" was introduced. This method is based on the minimum bunsetsu method and provides a practical technique for processing business documents in which a large number of prefixes and suffixes are used. In the "better method", dictionary searches and key input are performed simultaneously, which allows several candidates for a set of bunsetsus to be found. When their end character positions coincide for a set, an appropriate partitioning is judged to have been found and the candidates up to the position are evaluated. In evaluations, words are given a weight of 1, while prefixes and suffixes are weighted 0.5. The "lightest" candidate is given priority. When the end position of the clause does not match, priority is given to the longest matched bunsetsus.

All three of the methods described above are based on the longest match principle. The simplification they bring to keyboard operations has ensured their wide use today.

3. Human Interface in Kana-Kanji Conversion

As mentioned above, Japanese input technology has progressed from the character-by-character approach to the current word-by-word kana-kanji conversion approach.

The adoption of the kana-kanji conversion system has made it possible to use alphanumeric keyboards in Japanese-language processing. A CRT display was used as the soft copy terminal, and a Japanese text editor program was implemented on a microprocessor, allowing the first Japanese word processor to be developed.

The biggest problem standing in the way of kana-kanji conversion, however, was the existence of a large number of homonyms in the Japanese language [7].

The automation of homonym selection was therefore the biggest human interface problem in system development. In order to reduce the homonyms, morphological analysis was tested for bunsetsus. A further attempt to limit homonyms was made by analyzing usage frequency and adding temporal dictionaries.

Should these processes still not be able to arrive at the right homonym, the user was asked to select the correct choice on the display. This required an extra operation in which the keyboard was used to choose one homonym from among a number of choices. Reduction of these operations is one of the key points in Japanese-language keyboard input.

3.1 Homonym Reduction

There are many homonyms in Japanese. It comes from the fact that a number of Chinese characters have the same pronounciation, that is, a kana string corresponds to many Chinese characters. For example, a kana string "こう" (kou) corresponds to more than one hundred Chinese characters, like "交", "向" and "高". Another example, "しょう" (shou) also has more than two hundred homonyms, like "抄", "青" and "庄". Chinese characters usually cannot stand by themselves but need two or more characters to form a word. When two Chinese characters are combined to form a word, the number of homonyms is reduced. Even when these facts are true, more than thirty percent of Japanese words have several homonyms. For example, "こうしょう" (koushou) corresponds to twenty homonyms like "交渉", "公証" and "高尚". To reduce the number of homonyms is a problem in kana-to-kanji conversion. The following two methods have been found to be quite effective for reducing the number of homonyms and to promote efficiency in the selection of homonyms.

3.1.1 Reduction of Homonyms by Grammatical Analysis

Homonyms are classified into twelve groups according to the parts of speech. Verbs, auxiliary verbs and adjectives have special conjugational forms. Each conjugational form is accompanied by restricted parts of speech, special auxiliary verbs or particles. The indeclinable parts of speech in Japanese, such as nouns, proper nouns and numerals are also accompanied by restricted kind of particles. Thus morphological analysis of Kana string greatly reduces the number of homonyms.

For example, Kana string "こうしょう" (koushou) has twenty homonyms, however, if the string is accompanied by a conjugative part of adjective "な", the number of candidates is reduced to one, "高尚". In another case, the string is followed by Kana string "して" (shite), then the bunsetsu is analyzed to be a verb with an accompaning particle. The number of homonyms in this case reduces from twenty to four.

3.1.2 Improvement of Homonym Selection efficiency

Several homonyms may remain after morphological analysis. These homonyms are displayed on the CRT display screen, and the operator can select the right word from the displayed candidate homonyms. Candidate words are usually displayed one by one until the operator finds the correct word to input. The efficiency of selecting homonyms depends mostly on the order in

which the candidate words are displayed on the CRT screen. The Japanese word processor has the capability to count the frequency of word usage. This function is extremely helpful since the words used by every person or company vary considerably. For example, if a Japanese word processor is used in lawyer's office, then law terms will be frequently used. The word frequency counters count the number of times a word is used, and the most frequently used word among the homonyms is displayed first on the screen. This mechanism augments the probability of the right word among the homonyms being displayed on the screen first to more than 95%.

Another mechanism to improve the efficiency of homonyms selection is the homonym registers. When an operator is typing or editing a document, if a word is selected among homonyms, then the homonym registers change temporarily the order of the homonyms to display the selected word first on the screen. The registers memorize the temporal order of homonym words while the operator is processing the document. Because the same word will probably be used in the same document, the temporal order of homonyms helps to improve the efficiency of homonym selection and to speed up the input of Japanese sentences.

3.1.3 Homonym Reduction Using Word Co-occurrence Relations

Techniques for reducing homonyms by making use of information on the likelihood of words appearing together in the same sentence (co-occurrence relations) are now actively being studied. This method requires a co-occurrence relation table of word pairs. When a number of homonym candidates are generated in a sentence, word pairs appearing in the table are given priority.

In order to use word co-occurrence tables to evaluate all homonyms, however, it is necessary to prepare the ordering of a million or so word pairs. By classifying of words and making a co-occurrence relation table of classified groups, the homonyms can be reduced to a smaller sized co-occurrence table. For example, the 1,000 semantic categories found in a thesaurus were able to be used to classify all words, and the co-occurrence relations between these categories reduced homonyms to 1/7-1/10 [8]. This homonym reduction method has recently been commercialized for some Japanese word processors.

4. Proofreading and Formatting Support Systems

Knowledge engineering technology is now being employed in Japanese-language input systems to create proofreading and formatting support programs.

Natural language processing can be thought of in four different levels: phonemes, syntax, semantics and context. The kana-kanji conversion system we have been discussing consists of phoneme level processing, that is, a word-by-word system, and is known for being effective in the proofreading of mis-inputted text [9]. When expanded to levels of syntax and semantics, however, even more effective proofreading systems can be realized.

Not only spelling errors but also grammatical errors in the text can be detected. Not only are mistakes shown, the program gives concrete suggestions for correcting errors and styles. The UNIX "Writer's Workbench [10]" and "EPISTLE [11]" are the first systems of this kind.

The logical structure of a sentence also influences its format. In business letters, for instance, it is common for the address, text, date and other aspects to be placed according to a standard rule. If this standard rule was structured as a logical formula, the formatting of a document can be automated as a kind of expert system.

4.1 Proofreading Support Functions

The function for proofreading support consists of the following capabilities:

Detection of spelling errors
Detection of grammatical errors
Detection of inconsistency in expression

It is difficult for the person who is writing to detect errors in his or her own sentences, as he or she is not conscious of the mistakes. These errors are more easily identified by using the proofreading support function.

In order to detect spelling errors, sentences expressed in kanji are matched with the word dictionary so that unmatching parts can be pointed out as possible spelling errors.

Detection of grammatical errors can be performed by using syntax analysis. By parsing a sentence according to given Japanese syntax rules, grammatical errors are detected at parts where this parsing fail.

In Japanese-language expression, it is a general rule to consistently express the same word in kanji if it appears once in the same text. For this rule, the "KWIC" table for words appearing in a document is used to check the expressions of the same word in different ways. Those parts found to be inconsistent in terms of expression can then be pointed out to the user.

The proofreading function outlined above, however, does not achieve 100 percent error detection. Nevertheless, it is still very useful for the person creating sentences as support for proofreading work. In Japan, such proofreading systems have been introduced into newspaper companies and are supporting proofreaders.

4.2 Automatic Formatting Functions

Many documents are written under certain rules. For instance, scientific papers generally consist of a title at the top, followed by the author's name, abstract, chapters and sections, with insertions of illustrations as needed. Japanese letters usually begin with a seasonal greeting before going to the main topic.

If a system can extract the structure of a document

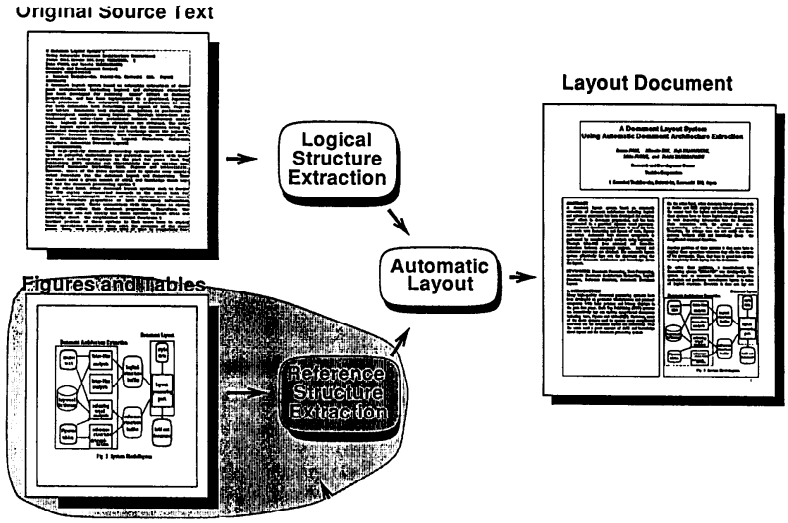


Fig. 3 Automatic Formatting System's outline.

and identify the components, such as the title, author's name and so on, the system can effectively lay out the document using the extracted document structure and knowledge about the layout [12]. The automatic formatting function, a kind of expert system using artificial intelligence, provides approximately 2,000 rules for extracting the text structure and also generates a layout format for scientific papers in Japanese.

When using the automatic formatting function in preparing a scientific paper, the average user can easily use this system without skill and knowledge about the complicated editing operations. Figure 3 shows an example of automatic formatting in which the original text is properly formatted. This system is now limited to scientific papers. However, the capability is now extending to business documents.

5. Conclusion

Most of today's Japanese inputting is done with the kana-kanji conversion system, which is found not only in word processors but as a front-end-processor or kernel in operating systems like MS-DOS and UNIX.

Though impossible in the 1960's, great strides have been made in Japanese-language computer processing in the '80's. Kana-kanji conversion has given all who desire it access to computers in Japanese.

Inputting Japanese will become even easier and faster in the future as more advanced natural language processing technology applied.

References

1. KAWADA, T. et al. Japanese Word Processor JW-10, *IEEE COMPCON FALL* (1979), 238–242.
2. KAWADA, T. et al. An Input Method of the Japanese Information, Institute of Electronics, *Information and Communication Engineers* (IEICE), EC-78-23, 1974.
3. KAWADA, T. Kana-chinese Translation, *IPS Japan*, **20**, 10 (Oct, 1979) 911–916.
4. MAKINO, H. Automatic Segmentation for Transformation of kana into kanji, *IPS Japan* (IPSJ), **20**, 4 (1979), 337–345.
5. YOSHIMURA, K. Morphological Analysis of Non-Marked-Off Japanese Sentences by the Least Bunsetsu's Number Method, *IPS Japan*, **24**, 1 (1983), 40–46.
6. SAITO, H., KAWADA, T. et al. Better input System, *IPS Japan* 1G-7 (1985).
7. KAWADA, T. Intellectual Word Processing and Its User Interface, *IEICE*, OS86-19 (1986), 1–6.
8. HONMA, S. et al. Translation of Non-Segmented Kana-Sentences to Kanji-Kana Sentences Using Collection Information, *IPS Japan*, **27**, (1986), 1062–1068.
9. KAWADA, T. Linguistic Error Correction of Japanese Sentences, *Proc. COLING-80* (1980), 257–261.
10. FRASE, L. T. The UNIX TM Writer's Workbench Software: *BSTJ*, **62**, 6 (1983), 1883–1890.
11. HEIDORN, G. E. The EPISTLE Text Critiquing system, *IBL SJ*, **21**, 3 (1982), 305–326.
12. IWAI, I. A Document Layout System Using Automatic Document Architecture Extraction, *Proc. CHI'89* (1989), 369–374.

(Received November 6, 1989)