JAPANESE WORD PROCESSOR

Shin-ya Amano
R & D Center, Information Systems Laboratory
Senior Researcher
Toshiba Corporation
1 Komukai - Toshiba-cho, Saiwai-ku, Kawasaki, 210

Japanese word processor (JWP) presents a typewriter which can easily handle Japanese documents with newly developed and quite different way from conventional Japanese typewriters. Japanese documents consist of more than two thousand letters (Kanjis: Chinese characters, Kanas: Japanese alphabet, alphanumerics, and so on). The conventional Japanese typewriter is equipped with all these letters. It means that typewriting is very difficult and typing speed is low. JWP overcame these difficulties with Kana-to-Kanji translation technology.

1. OVERVIEW

Japanese is a very unique language in its variety of letters. It has several kinds of letters: Hira-Kana (Japanese alphabet; あ, い, う, ---), Kata-Kana (Japanese alphabet used for foreign words; ア, イ, ウ, ---), Kanji (Chinese character; 山, 川, 東, ---), and alphanumerics (a, b, c, ---, 0, 1, 2, ---).

Kanas consist of about fifty letters. Popular Kanjis amount to about two thousand. These letters appear in ordinary Japanese documents such as newspapers, business letters, and so on. The abundant letters invoke difficulties in typing Japanese documents. The conventional Japanese typewriter has to be equipped with all letters, and typists must look for letters one by one. Thus "letter-searching" is an intrinsic problem in the Japanese typewriting.

In order to overcome this problem, Kana-to-Kanji tranlation technology is applied in the Japanese Word Processor (JWP). A Kana sentence is just a sequence of phonetic symbols. So the Kana sentences are easy to type because of the small number of keys, but

they are difficult to understand.

The new techniques to translate Kana sentences into ordinary Japanese sentences are discussed henceforth.

2. KANA-TO-KANJI TRANSLATION

Grammatical analysis is applied to get correct translation by JWP. Kana sentences are translated into Kanji sentences phrase by phrase.

The definition of the phrase is given as follows:

Phrase = (prefix) substantive word (suffix)
(functional word*)

Substantive word = noun/pronoun/verb/adjective/adverb/
conjunction/interjection

Functional word = particle/auxiliary verb

Here parentheses indicate optionality, the asterisk indicates one or more occurrences, and the slant indicates alternatives.

Generally speaking, nouns are accompanied with a particle which indicates cases. Verbs are accompanied with conjugating auxiliary verbs, which indicate tenses, moods, and voices, without any separation between them (Fig. 1).

みなければならなかった
minakerebanaranakatta.
(had to see)

Each underlined part denotes a verb,
auxiliary verbs, and a particle.

Fig. 1 An example of Japanese verb phrase

The grammatical analysis means analyzing the phrase grammatically. There are two main difficulties here, conjugation and concatenation of verbs and auxiliary verbs, and appearance of homonyms.

2.1 Grammatical Analysis

Japanese sentences consist of a series of the phrases. The role

of the Japanese phrases in sentences is approximately similar to English one (Fig.2), but structure of the phrases is different as shown by the above definition.

私は	学校へ	行く	.		I	go	to school.
p1	p2	p3			p1	p3	p2

(watasiwa gakkoue iku)

Here p1 is a noun phrase, p2 is a prepositional phrase, and p3 is a verb phrase.

Fig. 2 An example of the role of phrases

The analysis of verb phrase will be cited below since it is the most intricate. Japanese verbs do not express tenses, moods, and voices by their conjugation. If it is necessary to express them, auxiliary verbs which denote them respectively are concatenated to the verb. The verbs are classified into five conjugational groups and there are 25 auxiliary verbs. They have six or less conjugational forms. These numerous conjugational forms of all verbs and auxiliary verbs make the verb phrases very complex.

The conjugational forms indicate the condition of the concatenations of verb and auxiliary verbs in the phrase. That is, they indicate what kind of word is concatenated to the verb or auxiliary verb. We made a connection table for all verbs and auxiliary verbs to check the concatenation. Figure 3 shows examples of the structure of the verb phrases.

走る	(hasiru)	run
走った	(hasitta)	ran
走りたい	(hasiritai)	want to run

Fig. 3 Examples of the structure of the verb phrase

In the figure 3, each underlined part indicates a verb and an auxiliary verb.

2.1.1 Dictionary consultation

The first step of the grammatical analysis is to consult a dictionary for finding the substantive words. All possible candidates are extracted from the dictionary. We take the phrase "はしらない" (hasiranai; do not run) as an example to explain the grammatical analysis. Candidates for "はしらない" are shown by figure 4. Only stems are stored for verbs to reduce the number of items in the dictionary. The dictionary has the following contents for a word; Kanas, Kanjis, grammatical informations such as part of speech, conjugational type, etc., and frequency of use.

KANJI	KANA	PART OF SPEECH	(MEANING)
柱	はしら (hasira)	noun	(column)
橋	はし (hasi)	noun	(bridge)
箸	はし (hasi)	noun	(chopsticks)
走	はし (hasi)	stem of verb	(run)
葉	は (ha)	noun	(leaf)
歯	は (ha)	noun	(tooth)

Fig. 4 Candidates for "はしらない"

2.1.2 Analysis of conjugation and concatenation

The conjugational part of verb "走" (hasi) must be ら (ra), り (ri), る (ru), れ (re), or ろ (ro) according to the grammatical information of the dictionary. As the third Kana of はしらない is (ra), the verb "走" satisfies the conjugational condition. Then it must be checked that the conjugating form of "走ら" (hasira) can be accompanied with the string "ない" (nai). It is known by the connection table that the auxiliary verb "ない" can be attached to "走ら". "走らない" (hasiranai) satisfies the condition of the concatenation. Therefore "走らない" is one of possible translations. Nouns are accompanied with

specific particles. These particles are stored in a table. The attached Kana-string to "柱" (はしら ; hasira) is "ない" (nai). But this string is not found in the table, accordingly "柱ない" does not satisfy the condition of the concatenation as the phrase. By the same procedure, all the nouns of the candidates are also rejected. Thus only one translation "走らない" is got.

2.2 Homonyms

2.2.1 Reduction of homonyms by grammatical analysis

There are a lot of homonyms in Japanese. It results from the fact that plural Chinese characters have the identical pronunciation expressed by Kanas. Chinese character itself is a letter and not a word in Japanese. The words are composed of one or more Chinese characters. The words have homonyms, too. That is, plural words have the identical Kanas. JWP reduces the homonyms by the grammatical analysis. For example, the phrase "きょうりょくな" (kyouryokuna; powerful) is translated uniquely into "強力な" by analyzing it grammatically. If the grammatical analysis is omitted and only consultation of the dictionary is applied two homonyms occurred: "強力な" and "協力な". Here "協力な" is meaningless, since "協力" is a noun meaning "cooperation" and "な" (na) is a conjugating part of the adjective with which noun can not be accompanied. The meaningless translations like this often happen without the grammatical analysis.

In case of the simplest translation technique, a table of Kanjis and their corresponding Kanas is used (Fig. 5).

あ (a) ; 亜 , 亜 , 亜 , ---
あい (ai) ; 愛 , 相 , 哀 , ---
あか (aka) ; 赤 , 丹 , 朱 , ---
-
-
-

Fig. 5 Part of table of Kana-Kanji pairs

There are a lot of homonyms to a phonetic unit which is expressed by Kanas. For example, a phonetic unit "きょう" (kyou) corresponds to 137 Chinese characters; "協", "強", "京", and so on. "りょく" has six; "力", "緑", "朽", and so on. Moreover, " きょうりょく " has other possibilities of combination of phonetic units except "kyou-ryoku"; that is, "kyo-u-ryo-ku", "kyo-uryo-ku", and so forth. If including these, homonyms increase enormously and many meaningless translations occur.

2.2.2 Reduction of semantical homonyms by learning

JWP has two kinds of memories to avoid semantical homonyms: a short term memory and a long term memory. The former is valid during typing one document and is reset when other documents are typed. When homonyms appear on a CRT display of JWP and one of them is selected by the typist, the word is memorized by JWP. When the same homonyms occur, the memorized word is indicated on the CRT display as the most possible candidate word in the context of the text. The latter is implemented as an administrator of the frequency of word use: JWP administrates its dictionary regarding to the frequency. When one of homonyms is selected as mentioned above, the frequency of the word increases by one. Accordingly the dictionary is becoming suitable to a certain field such as science, economics, literature, etc., since JWP refers the frequency during translation and the most frequent word is given priviledge to appear on the CRT display first. This method is effective for the semantical homonyms depending fields.

3. CONCLUSION

The automatic Kana-to-Kanji translation made the touch method possible in Japanese typing and released the typists from the burden of searching letters among the huge number of letters on the keyboard.

JWP occupies a remarkable status against the conventional Japanese typewriter and other related machines for its ability of automatic translation. Several new techniques applied here are quite effective to realize high performance translation. Figure 6 shows an example of translation which is a Japanese sentence of the abstract of this paper.

にほんごワードプロセッサはこれまでのわぷんタイプライタとはまったくことなるあたらしくかいはつされたほうほうによりにほんごのぶんしょをよういにあつかえるタイプライタをていきょうします。

--- Typed Kana sentence ---

日本語ワードプロセッサはこれまでの和文タイプライタとは全く異なる新しく開発された方法により日本語の文書を容易にあつかえるタイプライタを提供します。

--- Translated sentence ---

Fig. 6 An example of Kana-to-Kanji translation

4. ACKNOWLEDGEMENT

We would like to thank Dr. Hiroshi Genchi and Dr. Ken-ichi Mori, our supervisors at Toshiba R&D Center for their encouragement and support in the course of the work reported here. We also wish to acknowledge Kimihito Takeda for his cooperation to us in developing JWP.