A Kana-Kanji Translation System for Non-Segmented Input Sentences Based on Syntactic and Semantic Analysis

Masahiro ABE, Yoshimitsu OOSHIMA, Katsuhiko YUURA and Nobuyuki TAKEICHI

Central Research Laboratory, Hitachi,.Ltd.
Kokubunji, Tokyo, Japan

Abstract

This paper presents a disambiguation approach for translating non-segmented-Kana into Kanji. The method consists of two steps. In the first step, an input sentence is analyzed morphologically and ambiguous morphemes are stored in a network form. In the second step, the best path, which is a string of morphemes, is selected by syntactic and semantic analysis based on case grammar. In order to avoid the combinatorial explosion of possible paths, the following heuristic search method is adopted. First, a path that contains the smallest number of weighted-morphemes is chosen as the quasi-best path by a best-first-search technique. Next, the restricted range of morphemes near the quasi-best path is extracted from the morpheme network to construct preferential paths.

An experimental system incorporating large dictionaries has been developed and evaluated. A translation accracy of 90.5% was obtained. This can be improved to about 95% by optimizing the dictionaries.

1. INTRODUCTION

Ordinary Japanese sentences are written using a combination of Kana, which are Japanese phonogramic characters, and Kanji, which are ideographic Chinese characters. Nouns, verbs and other independent words are generally written in Kanji. On the other hand, dependent words such as postpositons, and auxiliary verbs, etc., are written in Kana. While there are about fifty Kana, there are several thousand Kanji, thus making it difficult to input Japanese sentences into a computer system.

Extensive research has been carried out on methods of inputting Kanji in an attempt to realize rapid and easy input. Among the methods proposed, Kana-Kanji translation appears to be the most promising. In this method, input sentences are entered in Kana using a conventional typewriter keyboard, and those parts of the sentences which should be written in Kanji are translated into Kanji automatically. In this process a non-segmented input form is desirable for users because there is no custom of segmentation in writing Japanese sentences. Therefore, the ultimate goal of a Kana-Kanji translation scheme should be to achieve error-free translation from non-segmented Kana input sentences.

This paper describes a system for achieving high accuracy in the Kana-Kanji translation of non-segmented input kana sentences.

1.1 Disambiguation Approaches in Kana-Kanji Translation

If ambiguity were not a problem in non-segmented input Kana sentences, a perfect Kana-Kanji translation could be easily made using simple transliteration techniques. The fact is that the input Kana sentences are highly ambiguous. The ambiguity of non-segmented input Kana sentences can be categorized into following two types.

(a) The ambiguity of segmentation of a sentence into morphemes.

example

(Input Kana sentence)		(Output sentence)
① ココデハ/イル。	→	ここでは要る。
[kokodeha iru]		(It's necessary here.)
② ココデ/ハイル。	→	ここで入る。
[kokode hairu]		(Enter here.)

(b) The ambiguity of homonyms.

example

(Kana)		(Homonyms)
キシャ	→	① 汽車 (a train)
[kisha]		② 貴社 (your company)
		③ 記者 (a pressman)
		④ 帰社 (return to office)
		⑤ 喜捨 (donate)

Makino and Kizawa [1] proposed an automatic Kana-Kanji translation system in which these two types of ambiguity are treated separately in different ways: The segmentation of input sentences is carried out heuristically by the longest string-matching method of two "Bunsetsu". A Bunsetsu is a Japanese syntactic unit which usually consists of an independent word followed by a sequence of dependent words. After determining the segmentation of a sentence, suitable words are selected from the homonym set based on a syntactic and semantic analysis. In their approach, the ambiguity of the segmentation is treated without using syntactic and semantic analysis.

The new Kana-Kanji translation method presented in this paper treats both types of ambiguity in the same

way based on a syntactic and semantic analysis. In the new method, translation is performed in two steps. In the first step, the both kinds of ambiguity are detected by morphological analysis and are stored in a network form. In the second step the best path, which is a string of morphemes, is chosen from the network by syntactic and semantic analysis based on the case grammar.

2. EXTRACTION OF AMBIGUITY FROM INPUT KANA SENTENCES

This section first describes the method for extracting highly possible ambiguities by morphological analysis, and then describes an efficient data structure for storing those ambiguities in memory.

2.1 Morphological Analysis

2.1.1 Morphological characteristics of Japanese language

A Japanese sentence is composed of a string of Bunsetsu, and each Bunsetsu is a string of morphemes. In a Bunsetsu the relationship between the preceeding morpheme and succeeding morpheme is strongly regulated by grammar. The grammatical connectability between morphemes can be easily determined by using a grammatical table in morphological analysis [2]. On the other hand, on the morphological level there is little if any grammatical restriction between the last morpheme in a Bunsetsu and the first morpheme in the following Bunsetsu. In this sence a compound word is also a series of Bunsetsu, each of which contains an independent word. There is no limit to the length of a compound word, and there are no restrictions in the way words can be combined. Therefore, since there are a tremendous number of compound words, it is almost impossible to compile a dictionary of these words.

2.1.2 Morpheme chain model

The lack of restrictions on the relationship of consecutive Bunsetsu increases the ambiguity of segmentation in the morphological analysis. This is especailly true if the formation of compound words is not restricted in some way. Under these circumstances the result is often meaningless because compound words are generated mechanically.

This problem can be solved by introducing the concept of a statistical model of a morpheme chain. Statistical research in this area [3] indicates that compound words have some distinct morphological characteristics:

(1) *Part of speech:* about 90% of morphemes in compound words are nouns or their prefixes or suffixes.

(2) *Word category:* about 77% of all morphemes are words of foreign origin (Chinese).

(3) *Length:* About 93% of compound words are 3 to 5 Kanji in length.

These properties can be used to distinguish very likely candidates for compound words from unlikely ones. Morpheme M can be represented by the property set (P, C, L), where P, C and L mean the part of speech, the word category and the length in Kana, respectively. A compound word can then be modeled as a morpheme chain, and is represented by pairs of the property set. The pairs can be classified into three levels according to the probability of occurrance. To generalize the representation a null property set $(-,-,-)$ is introduced for the edge of a compound word. Table 1 is a part of the model representation.

2.1.3 Algorithm of morphological analysis

Figure 1 shows the algorithm for the morphological analysis. All candidates for dependent words are first picked up from input Kana sentences by using a string-matching operation and by examining the grammatical connectability between a preceding word and its successor. This process is executed from right to left, resulting in the generation of subtrees of dependent words.

Next, candidates for independent word are picked up by string-matching using a word dictionary starting from the leftmost character of the input sentence. Those

Table 1 Statistical Model of Morpheme Chain

Level	*M1*			*M2*		
	P	*C*	*L*	*P*	*C*	*L*
1	–	–	–	any	any	(*1)
1	noun	Chinese	≧3	noun	Chinese	≧3
1	noun	Chinese	≧3	suffix	Chinese	2
1	noun	Japanese	≧2	noun	Chinese	≧3
:	:	:	:	:	:	:
2	noun	Japanese	2	noun	Japanese	2
2	noun	Chinese	2	suffix	Chinese	2
:	:	:	:	:	:	:
3	noun	Chinese	≧3	unknown	any	any
:	:	:	:	:	:	:

(*1): longest or 2nd longest matching words
M1: Preceeding morpheme
M2: Succeeding morpheme
P: a part of speech
C: word category
L: length in Kana

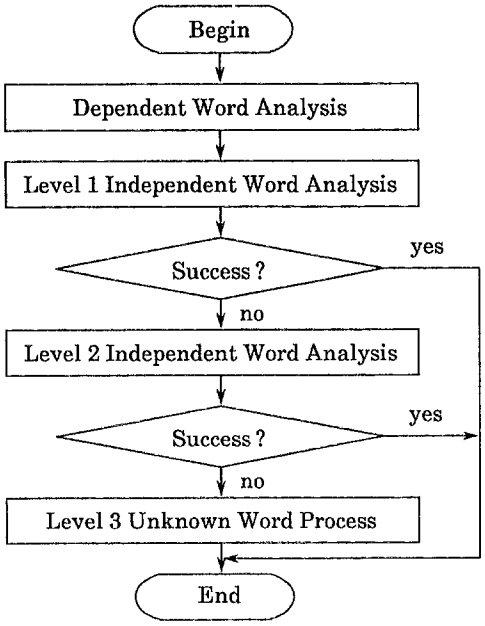

Figure 1 Algorithm for Morphological Analysis

elements which correspond to the level 1 chain are then selected. If the selected independent word adjoins a dependent word which has already been extracted in the previous process, the grammatical connectability between them is also checked. In this way all independent level 1 words that begin from the first column are extracted. The control is shifted to the next column of the input sentence. If the current position is the ending edge of the already extracted independent word or its successive dependent word, the same action is taken. If not, the control moves to the next column without extracting any independent words. The control stops when it reaches the end of the sentence after having successfully extracted all level 1 independent words or related successive dependent words.

If the system fails to extract any words on level 1, the control backtracks to the beginning, and trys again using level 2 extraction. On this pass, level 2 independent words are picked up and tested in the same manner as in level 1 extraction. If the level 2 extraction fails, then an unknown word process, level 3, is invoked, which assumes an unknown word exists in the input sentence and the control skips to the nearest dependent word. The skipped part is assumed to be the unknown word. In this way, the control of the extraction level for independent words based on a statistical model of morpheme chains enables highly possible ambiguities in input Kana sentences to be extracted by pruning rare compound words.

2.2 Network Representation of Ambiguity

The ambiguous morphemes extracted in the morphological analysis are stored as common data in a network form to reduce both storage and processing overhead. Figure 2 shows an example of a morpheme network. Each morpheme is represented by an arc. Both ends of each morpheme are indicated by circles. A double circle corresponds to the end of a Bunsetsu, whereas a single circle corresponds to the boundary of a morpheme in a Bunsetsu.

The information for a group of ambiguous morphemes is represented by the data structure: *VTX* (Vertex), *EDG* (Edge) and *ABL* (AmBiguity List). The *VTX* represents the position of morpheme in a sentence. The *EDG* represents the common attributes of the ambiguous morphemes. The common attributes are a part of speech, the type of inflection and Kana string. The *ABL* represents individual attributes of the morphemes. The individual attributes are the Kanji representation, the meaning code and the word frequency. An *ABL* list is referenced by *EDG*. *VTX* and *EDG* refer to each other. A *VTX* is considered to be shared if the grammatical relationship between the preceeding *EDG* and its succeeding *EDG* is the same. A double circled *VTX* can usually be shared.

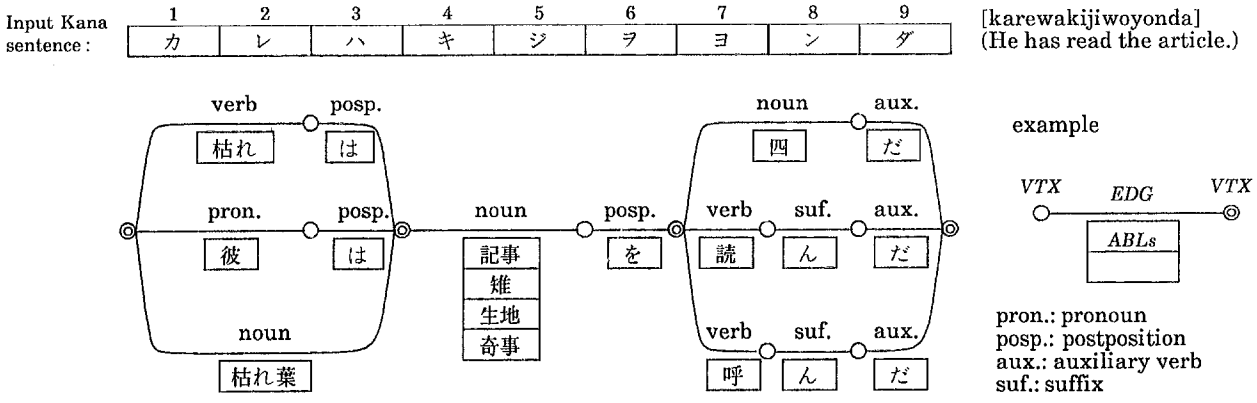

Figure 2 An Example of a Morpheme Network

3. SELECTION OF THE MOST SUITABLE MORPHEME STRING

The second step in the Kana-Kanji translation process is divided into two substeps:

(1) Extraction of morpheme strings from the morpheme network.

(2) Selection of the best morpheme string by syntactic and semantic analysis.

3.1 Extraction of Preferential Paths

Each path, or morpheme string, can be derived by tracing morphemes on the network from the beginning of the sentence to the end of the sentence. In order to avoid the combinatorial explosion of possible paths, it is necessary to introduce some heuristics which make it possible to derive highly possible paths. This is accomplished in the following way. First, a quasi-best path is chosen based on the least weighted-morpheme number using the best-first-search technique [4]. Next, a restricted range of morphemes near the quasi-best path is selected from the morpheme network in light of the probability of ambiguity.

3.1.1 Least weighted-morpheme number

The least Bunsetsu number [5] is known as an effective heuristic approach for determining the segmentation of non-segmented input Kana sentences. In this approach, the segmentation which contains the least number of Bunsetsu is most likely to be correct. The authors have modified this method to improve the correctness of segmentation by changing the counting unit from the number of Bunsetsu to the sum of the weighted morphemes. The weights of morphemes are basically defined as 1 for each independent word and 0 for each dependent word. Since a Bunsetsu is usually composed of an independent word and some dependent words, the sum of the weights of a sentence is roughly equal to the number of Bunsetsu in the sentence. While the least bunsetsu number ignores the contents of the Bunsetsu, the new method evaluates the components of the Bunsetsu to achieve more realistic segmentation. The weights morphemes were modified based on empirical statistical data. Consequently, some independent words such as Keishikimeisi (a kind of noun), Hojodoshi (a kind of verb) and Rentaishi (a kind of particle) are weighted 0.1. The weight of prefixes and suffixes in compound words is defined 0.5. Table 2 shows the weight for morphemes.

Table 2 Morpheme Weighting

part of speech	weight
Nouns, Verbs, Adjective Verbs, Adverbs, Conjuctions, Interjections	1.0
Prefixes, Suffixes	0.5
Keishikimeishi, Hojodoushi, Rentaishi	0.1
Others	0

3.1.2 Best-first-search for a quasi-best path

In Figure 3, *VTX(0)* and *VTX(n)* correspond to the beginning and the end of a sentence, respectively. Each *VTX* and each *EDG* contains a cumulative number of weighted morphemes beginning from the end of the sentence. They are represented *W(i)* for *VTX(i)* and *W(ij)* for *EDG(ij)*. *X(ij)* is the weight of the *EDG(ij)*.

For the *VTX(n)*

$$W(n)=0 \quad \text{----------(1)}$$

Generally, for *EDG(ij)*

$$W(ij)=W(j)+X(ij) \quad \text{----------(2)}$$

And for *VTX(i)*

$$W(i)=\min_{j}\{W(ij)\} \quad \text{----------(3)}$$

This means that the minimun *W(ij)* is selected among the *EDG*s which share *VTX(i)* on their left side. By repeating (2) and (3), the minimum sum of the weighted-morpheme number can be got as *W(0)*. Then a quasi-best path which has a least weighted-morpheme number can be easily obtained by tracing the minimum *W(ij)* starting from the *VTX(0)*. Since the complexity of the above process is on an oder of n, the quasi-best path can be obtained very efficiently.

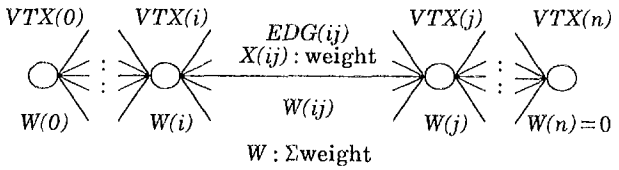

Figure 3 Best-first-search on Morpheme Network

3.1.3 Selection of alternative paths

Since the selected quasi-best path is not always the most suitable one, alternative paths are created near the quasi-best path by combining the restricted range of ambiguous morphemes. The range is decided by a preferential ambiguity relationship table (See Table 3) which contains typical patterns of segmentation ambiguity. By looking up this table, highly possible ambiguities for morphemes of the quasi-best path can be selected from the morpheme network.

3.2 Syntactic and Semantic Analysis

3.2.1 A meaning system

A detailed general purpose meaning system is necessary for Kana-Kanji translation. The meaning system adopted was basically a thesaurus of 32,600 words

Table 3 Preferential Ambiguity Relation

quasi best path	The range of alternative ambiguous morphemes					
	n.	v.	a.	a.v.	adv.	posp.
n.		○		○		○
v.	○	○			○	○
a.						○
a.v.	○					○
adv.		○				○
posp.	○	○	○	○	○	○

n.: nouns, v.: verbs, a.: adjectives, a.v.: adjective verbs, adv.: adverbs, posp.: postpositions

classified into 798 categories of 5 levels [6]. The system was enhanced by adding 11 higher level meaning codes called macro codes, such as "human", "thing" and "place". Each macro code was made by gathering related meaning codes in the system. In the original system, these codes appeared in different categories. The word dictionary developed for the new system contains 160,000 words. Each word is given a meaning code according to the new meaning system.

3.2.2 Case grammar and case frames

Case grammar [7] is widely used in natural language processing systems. It is also useful in Kana-Kanji translation because it can be applied to homonym selection as well as to syntactic analysis. When used for this purpose, the case frame must have a high resolving power so that it can distinguish a correct sentence from among many ambiguous sentences. The way in which the new approach achieves high resolving power in case frames can be summerized as follows:

(1) Detailed meaning description in case frames.
Each slot in a case frame has a list of meaning codes that fit for each case. The meaning codes are written in the lowest level of the meaning system except when higher meaning codes are preferable. In special cases, such as when an ideomatic expression is required for a case slot, a word itself is written instead of the meaning code.

(2) Rank specification of cases.
Cases are classified into either obligatory or optional cases.

(3) Multi-case frames for each verb.
A case frame is provided corresponding to each usage of a verb.

A case frame dictionary of 4,600 verbs was developed for this system.

Table 4 shows an example of case frame description. Each case frame consists of case slots and information about the transformation such as voice. Each case slot contains the case name, the typical postposition associated with the surface structure, the case rank indicator and meaning codes.

3.2.3 Parsing algorithm

Syntactic and semantic analysis is performed concurrently. Moreover, the homonym selection is made simultaneously. The process is basically a pattern-matching of paths with the case frame dictionary and is performed as follows. A path is scanned from left to right. Every noun Bunsetsu which depends on a verb in the path is pushed down to a stack. Whenever a verb is encountered during scanning, case frame matching is carried out. Every combination of noun Bunsetsu and case slots of the verb are tried and evaluated. The best combination is determined using the following conditions:

(1) Coincidence of postpositions.
The postposition of the noun Bunsetsu must be equal to the one for the case slot.

(2) Coincidence of meaning code.
The meaning code of the noun must be equal to the one for the case slot. If the noun has homonyms in ABL, a coincident homonym is selected.

(3) Occupation of obligatory case slots.
Higher occupation of obligatory case slots is preferable.

Table 4 An Example of a Case Frame

Case Frame	Case Slots				Type of Transformation
	Name	Rank	Postposition	Meaning Code	
読む#1 (read #1)	agent	obligatory	が [ga]	5.11	使役変形#4 (Causative #4) 受身変形#3 (Passive #3)
	object	obligatory	を [wo]	1.31, 1.320, 1.321, 1.3080, 1.19, 1.17, 1.309303	
	place	optional	で [de]	5.3	
	time	optional	に [ni]	5.2	
	---	---	---	---	
---	---	---	---	---	---

(4) Total occupation of case slots.

To addition to the condition (3), higher total occupation of case slots is preferable.

If using the above conditions it is not possible to choose a single combination, then word frequency information is used. throughout this process, unmatched noun Bunsetsu are left in the stack and are assumed to depend on verbs which occur later in the path. This case frame matching is repeated everytime a verb is encountered in the path. The parsing result of the path is obtained when the scanning reaches the end of the path. The same parsing is tried for other paths constructed in the previous step. Then the most suitable path is selected among the successfully parsed paths by measuring the degree of fit for conditions (3) and (4) above. The result is the text of the Kana-Kanji translation.

4. EXPERIMENT

4.1 System Implementation

The experimental system developed by the authors is shown in Figure 4. The system consists of three subsystems: a translation control program, a morphological analysis program and a syntactic and semantic analysis program. The total size of the system is about 35K steps in PL/I. Two large dictionaries are also developed: a word dictionary of 160,000 entries and a case frame dictionary of 4,600 verbs.

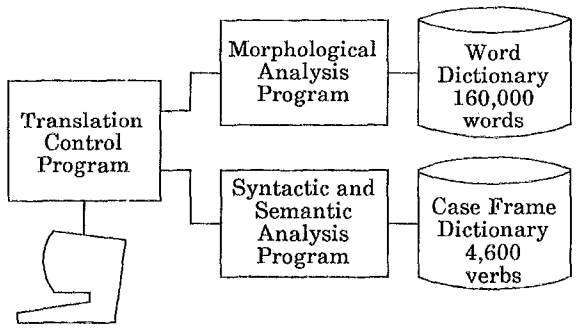

Figure 4 Kana-Kanji Translation System

4.2 Experimental Results

Both the upper limit and the substantial level of the accuracy of the new Kana-Kanji translation system was determined experimentally. The upper limit of translation accuracy was determined using a set of benchmark texts consisting of 9 typical Japanese texts including official documents, scientific papers and legal documents. The total number of input characters in the benchmark text was about ten thousand. Program errors and data errors in dictionaries were corrected to as great an extent as possible. The accuracy of the system using the benchmark texts was 94.9%. Another set of Japanese texts was prepared with twenty thousand chracters and the translation exreriment was repeated. This time no correction of data errors was made during the experiment. The average accuracy was 90.5%, which is the current level of performance of our system.

5. CONCLUSION

A new disambiguation approach to Kana-kanji translation for non-segmented input sentences has been described. Ambiguitiy is resolved using syntactic and semantic analysis based on a case grammar. To avoid a combinatorial explosion of the ambiguity, some heuristics are introduced. Large dictionaries were also developed for the experimental system and both the limit and substantial performance of the system were evaluated. The experimental results show that an accuracy of 90.5% is obtainable using this approach, and that the accuracy can be improved to about 95% by optimizing the dictionaries. Further improvement can be achieved by introducing context analysis and plagmatic analysis.

Acknowledgements

The authors wish to thank Dr. Mutsuko Kimura, Senior Researcher, Institute of Behavioral Sciences for her help in making the case frame dictionary. The authors are also indebted to Professor Makoto Nagao, Department of Electrical Engineering, Kyoto University for his thoughtful comments and to Dr. Hisashi Horikoshi, Manager, Planning Office, Central Research Laboratory, Hitachi,Ltd., for his constant encouragement throughout the course of this work.

REFERENCES

[1] M.Makino and M.Kizawa, "An Automatic Translation System of Non-segmented Kana Sentences into Kanji-Kana Sentences", COLING80, pp. 295-302 (1980).

[2] I.Aizawa and T.Ebara, "Machine Translation System of 'Kana' presentations to 'Kanji-Kana' Mixed prentations", NHK. Tech. Res., pp. 261-298 (1973).

[3] O.Hayashi, editor, "Statistical Charts of Japanese Language", Kadokawa Shoten, Tokyo (1982).

[4] A.Barr, E.A.Feigenbaum, "The Handbook of Artificial Intelligence Pitman, Vol.1, pp. 58-63 (1981).

[5] K.Yoshimura, T.Hitaka and S.Yoshida, "Morphological Analysis of Non-marked-off Japanese Sentences by the Least BUNSETSU's Number Method", Johoshori, Vol 24. No.1, pp. 40-46 (1983).

[6] National Language Research Institue, "Word List by Semantic Principles", Syuei Syuppan, Tokyo (1964).

[7] C.J.Fillmore, "The Case for Case", in Universals in Linguistic Theory. Edited by Emmon Bach and Robert T. Harms, pp.1-90, Holt, Rinehart and Winston, Chicago(1968).