JAPANESE SENTENCE ANALYSIS FOR AUTOMATIC INDEXING

Hiroshi Kinukawa

Systems Development Laboratory
Hitachi, Ltd.
1099, Ohzenji, Tama-ku,
Kawasaki 215, Japan

Hiroshi Matsuoka

Systems Development Laboratory
Hitachi, Ltd.
1099, Ohzenji, Tama-ku,
Kawasaki 215, Japan

Mutsuko Kimura

Institute of Behavioral Sciences

1-35-7, Yoyogi, Shibuya-ku,
Tokyo 151, Japan

A new method for automatic keyword extracting and "role" setting is proposed based on the Japanese sentence structure analysis. The analysis takes into account the following features of Japanese sentences, i.e., the structure of a sentence is determined by the noun-predicate verb dependency, and the case indicating words (kaku-joshi) play an important role in deep case structure. By utilizing the meaning of a noun as it depends on each predicate verb, restricted semantic processing becomes possible. An automatic indexing system, equipped with a man-machine interactive error-correcting function, has been developed. The evaluation of the system is performed by applying it in news information retrieval. The results of this evaluation show that the system can be put to practical use.

1. Introduction

The main problems arising with the development of an information retrieval system for the Japanese text are the need for saving man-power, standardizing information storage, and the realization of efficient retrieval. In the case of the English text, the stop-word removing method for automatic keyword extraction has been put to practical use. However, in the case of the Japanese text which consists of Kanji and Kana characters, a keyword extraction method utilizing statistical word frequency data has been reported by a Kyoto University group.[3] This paper proposes a new method of automatic keyword extraction and "role" setting for Japanese news information retrieval. The "role" characterizes semantic identification of each keyword in a sentence and is classified into six categories, i.e., human subject, human object, time, place, action, and miscellaneous important information.
The main features of Japanese sentences can be characterized as follows:

(1) The structure of a sentence is determined by the noun-predicate verb dependency.
(2) The case indicating words(kaku-joshi) play an important role in deep case structure.

Taking these features into account, D.G.Hays's dependency grammar[1] and C.J.Fillmore's case grammar[2] are utilized in the sentence structure analysis. The sentence pattern table containing a noun-predicate verb dependency relationship plays an important function in the analysis. By utilizing the meaning of a noun as it depends on each predicate verb, restricted semantic processing becomes possible. An automatic indexing system[5], equipped with a man-machine interactive error-correcting function, has been developed based on the method described. Evaluation of the system has been done by applying it in news information retrieval.

2. Role Setting Criteria

The employed criteria for the role setting of each keyword in a news sentence are as follows:

(1) "Action"(A for short) is assigned to verbs which express movement and are elements of the "predicate" set.
(2) "Time"(T for short) can be assigned without ambiguity.
(3) "Human subject"(HS for short), "human object"(HO for short), "place"(P for short) and "miscellaneous important information" (MI for short) are assigned to noun words according to the following criteria:
 (a) Words which express humans or organizations have either role "HS" or "HO". The distinction can be made by examining the subsequent kaku-joshi.
 (b) Words which express things without consciousness have role "MI".
 (c) A country name has role "HS" if it is presumed to have consciousness as an organization. It has role "P" if it means territory.
 (d) An airplane or a ship have role "HS" when they are personified together with the driver, role "P" when they express the place, and role "MI" when they mean things.
 (e) Ambiguities in item (c) and (d) are removed by knowing which predicate verb the word depends on and this determine which human, organization, place or miscellaneous matter it expresses.

To clarify the description, some examples are given below:

ex.1) "State A"ga "State B"wo shihai-suru.
HS HO A (control)
<State A controls State B.>
In this sentence "ga" and "wo" are kaku-joshis.

ex.2) "State A"ga "M-Sea"wo shihai-suru.
HS P A
<State A controls M-Sea.>

ex.3) "State A"ga "petroleum"wo
HS MI
shihai-suru.
A
<State A controls petroleum.>

ex.4) "Isolationism"ga "State A"wo
MI HO
shihai-suru.
A
<Islationism controls State A.>

(4) As mentioned above the "role" of a noun word is determined by considering the following three elements: i.e.,
(a) the predicate verb which the noun word depends on
(b) the meaning of the noun word
(c) the kaku-joshi which is concatenated to the noun word

3. Japanese Sentence Structure Analysis

The basic Japanese sentence pattern is expressed as "NF_1NF_2--NF_nPV", where NF_i, which is called "meishi-bunsetsu", is composed of a noun word and case indicating words, and where PV is a predicate verb.
The Japanese sentence structure is characterized by the following points. i.e.,
(1) The predicate verb is put at the end of the sentence.
(2) The position of a "meishi-bunsetsu" in a sentence is not fixed.
(3) A "meishi-bunsetsu" could be omitted in a discourse which consists of several sentences.
Utilizing D.G.Hays's dependency grammar, noun-predicate verb dependency relationships are formulated. In this formulation the relationships between nouns are irrelevant. Therefore, the Japanese sentence structure becomes independent of noun-word order, and a word omission is expressed in terms of the presence of a dependency relationship in the sentence. Since "role" is semantic identification of a word, by applying C.J.Fillmore's case grammar[2], it can be assigned to each keyword by clarifying the case structure of the predicate verb.(Figure 1) In Japanese sentence structure analysis, the predicate verb is identified first and then dependent noun words are determined in order of nearness to the predicate verb. The sentence is parsed by using top-down analysis. The bottom-up method is not adopted because it causes much ambiguity in the parsing of words which do not directly depend on the predicate verb. The need for classification of noun words in terms of their meaning is mentioned in chapter 2. Noun words are classified into seven semantic classes in order to analyze noun-predicate verb dependency relationships efficiently and to set "role"s to them, i.e.,
(i) Organization (ii) Person
(iii) Literature (iv) Place (v) Action
(vi) Name of matter, Abstract idea, etc.
(vii) Time
Predicate verbs are classified by taking into account the meaning of the dominated words and their cases. (Figure 2) The sentence pattern table is constructed based on this predicate verb classification. (Figure 3)
In the news retrieval system, about 5600 predicate verbs are classified into 586 classes; this classification is called case-information(A4-code). The sentence pattern table contains 1686 patterns. A Sentence pattern in the table is composed of four triplets at most. Elements of the triplet are the semantic class identification code of the noun word, kaku-joshi, and the "role" which is determined in terms of the values of the first two elements.
For example "shihai-suru"(control) and "kogeki-suru"(attack) belong to No.46 category. The predicate verb of this category has six sentence patterns and each sentence pattern has two triplets. The first sentence pattern has triplets (ga,A,1) and (wo,1,2). The first code of the triplet is "kaku-joshi", the second

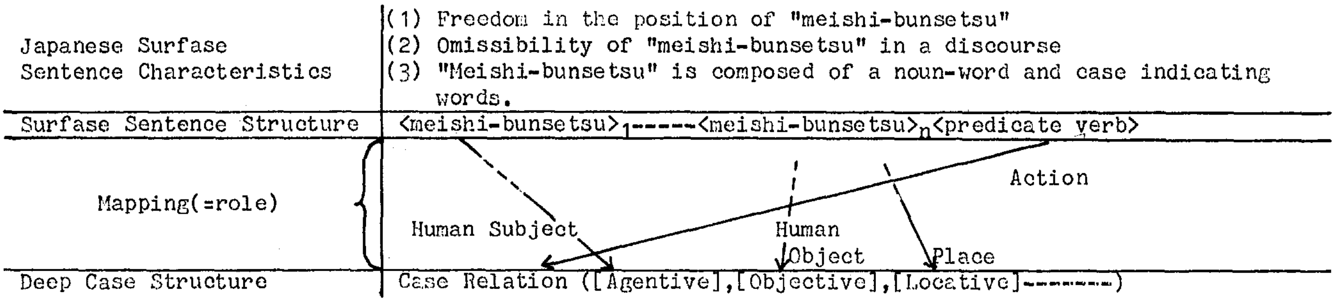

Japanese Surfase Sentence Characteristics
(1) Freedom in the position of "meishi-bunsetsu"
(2) Omissibility of "meishi-bunsetsu" in a discourse
(3) "Meishi-bunsetsu" is composed of a noun-word and case indicating words.
Surfase Sentence Structure
<meishi-bunsetsu>$_1$-----<meishi-bunsetsu>$_n$<predicate verb>
Mapping(=role)
Human Subject
Human Object
Place
Action
Deep Case Structure
Case Relation ([Agentive],[Objective],[Locative]-------)

Figure 1 Relationship between Surfase Case Structure and Deep Case Structure in Japanese Sentence

code is the semantic classification code of the noun word, and the third code is the "role". Semantic classification code "A" expresses organization or person.
Sentence analysis and "role" setting are performed referring to this sentence pattern table.

4. Automatic Indexing System

An automatic indexing system has been developed based on the method described. The processing procedure of the system consists of the following three steps(Figure 4):
(1) Word recognition

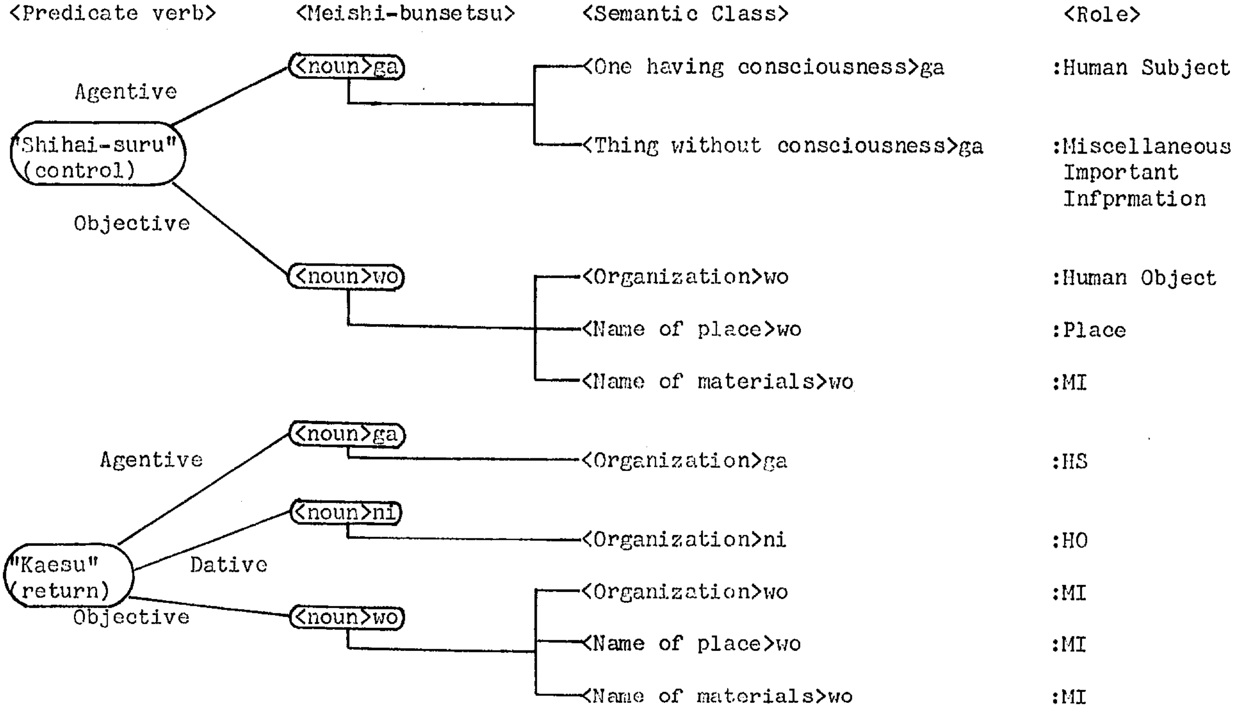


Figure 2 Relationship between predicate verb and roles

A4code	First			Second			Third			Fourth		
	K	B	R	K	B	R	K	B	R	K	B	R
1	ga	1	1									
≀												
46	ga	A	1	wo	1	2						
Shihai-suru	ga	A	1	wo	4	4						
(control)	ga	A	1	wo	6	6						
Kogeki-suru	ga	6	6	wo	1	2						
(attack)	ga	6	6	wo	4	4						
etc.	ga	6	6	wo	6	6						
≀												
586												

K:Kaku-joshi
B:Semantic Identification of Noun Words
1:Organization 2:Person 3:Literature
4:Place 5:Action 6:Name of materials,etc.
7:Time A:1 or 2
R:Role
1:Human Subject 2:Human Object
3:Time 4:Place
5:Action 6:Miscellaneous Important Information

Figure 3 Proposed Sentence Pattern Table

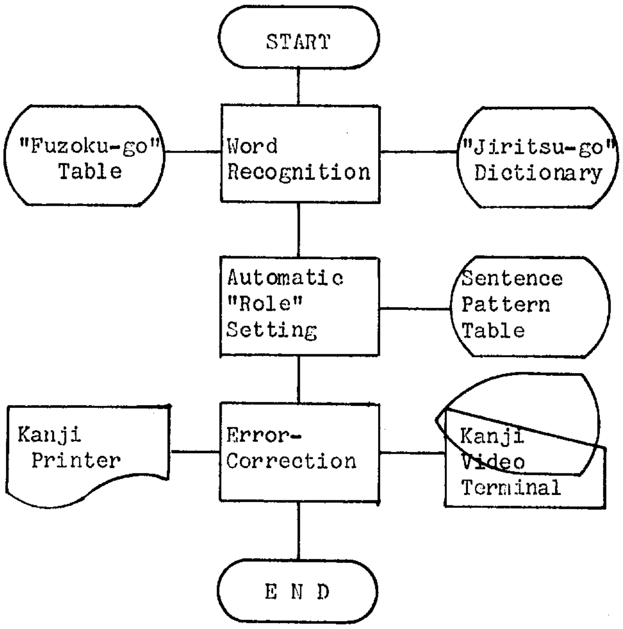


Figure 4 Automatic Indexing Procedure

(2) An automatic "role" setting resulting from the sentence structure analysis
(3) Man-machine interactive error-correction.

The hardware configuration is given in Table 1. Size and performance of the programs are given in Table 2.

4.1 Word Recognition

Word recognition is executed in the following two steps,(Figure 5) i.e., automatic segmentation of the Kanji and Kana character string, and the matching of each segment with entries in the content word dictionary ("Jiritsu-go" dictionary which contains nouns, verbs, etc.) and the function-word table ("Fuzoku-go" table) to obtain syntactic and semantic information concerning the word. The first step utilizies statistical features of Japanese sentences. The second step is a morphological word analysis[4]. The following information codes are given to the words contained in the "Jiritsu-go" dictionary:

(1) A1-code:ten word-class classification code
(2) A2-code:75 morphological class classification code
(3) A3-code:prefix and suffix identification code
(4) A4-code:predicate-verb case identification code
(5) B-code :semantic identification of noun words

The morphological analysis procedure gives the following information by referring to the "Fuzoku-go" table:

(6) C1-code:kaku-joshi classification code
(7) C2-code:the code distinguishes active voice, passive voice and causative expression
(8) C3-code:The code given to a meishi-bunsetsu distinguishes whether the meishi-bunsetsu is a direct dependant of the predicate-verb or a modifier of another meishi-bunsetsu.
The code given to the prdicate-verb expresses the type of inflection of the verb and the kind of subsequent conjunctive function word(setsuzoku-joshi).
(9) D-code :auxiliary code for determining A1-code

Table 1 Hardware Configuration

No.	Name	Specification and Usage
1	C.P.U.	Memory:384KB S.M.V.:3.6 s.
2	M.Disk	M.A.T:72.5ms.
3	Kanji Printer	Dictionary & Table str. media 700line/min. Printing of results
4	Kanji Video Terminal	40ch./line x 12 line Man-machine interactive error-correction

C.P.U.:central processing unit
M.Disk:magnetic disk memory
S.M.V.:system mixed value
M.A.T.:mean access time
str. :storage
min. :minute
ms. :milli-second

Table 2 Size and Performance of the Programs

No	Procedure	Steps	Memory	Pfm.
1	Word Recognition	3 KS	60KB	240ms/m.b.
2	Automatic "Role" Setting	12	120	650ms/stc.
3	Error-Correcting	6	132	—
4	Table Maintenance	6	33	—
5	Utility	11	84	—
6	Total	38	132	

These procedures are programmed in Assembly language.
KS :kilo-steps
KB :kilo-byte
m.b.:meishi-bunsetsu
Pfm.:performance
stc.:sentence

4.2 Automatic "Role" Setting

Automatic "role" setting is executed by the following four steps(Figure 6):

(1) Predicate verbs in a sentence are recognized by referring to the A1-code at first. Then, complex sentence structure is analyzed and divided into simple sentences.
(2) Sentence patterns for each simple sentence are obtained by utilizing the A4-code. Then, noun-predicate verb dependency is analyzed by comparing the B-code and the C1-code of noun words with the sentence pattern. Prior to this analysis the following procedures are executed.
(a) Seaching the sentence pattern for causative expression
(b) Transforming passive voice expression into active voice expression
(c) Standardizing "kaku-joshi"

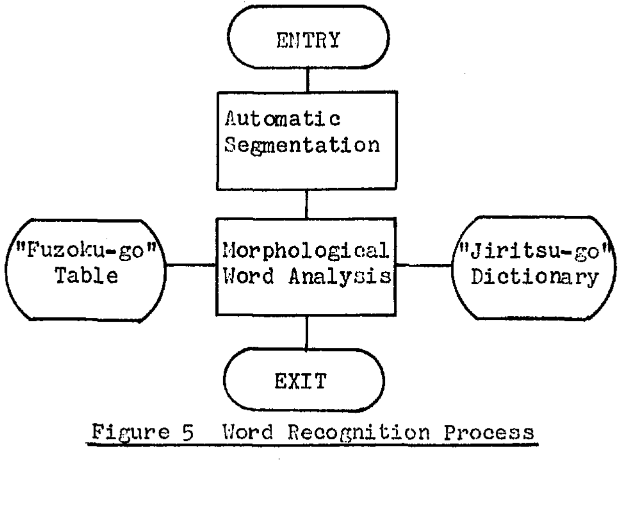


Figure 5 Word Recognition Process

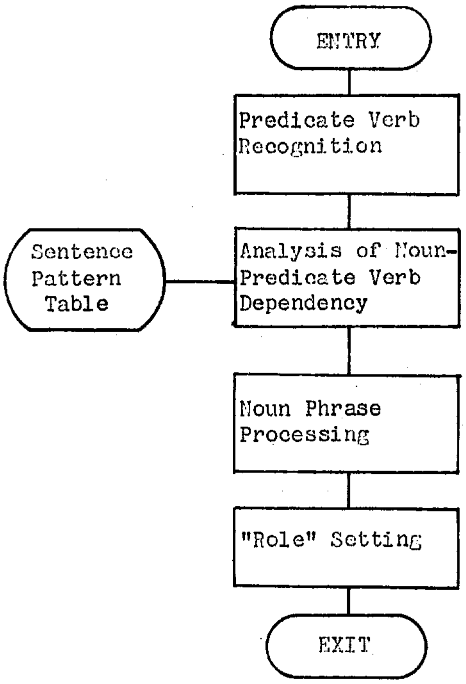

ENTRY
Predicate Verb Recognition
Sentence Pattern Table
Analysis of Noun-Predicate Verb Dependency
Noun Phrase Processing
"Role" Setting
EXIT

Figure 6 Role-Setting Process Utilizing Japanese Sentence Analysis

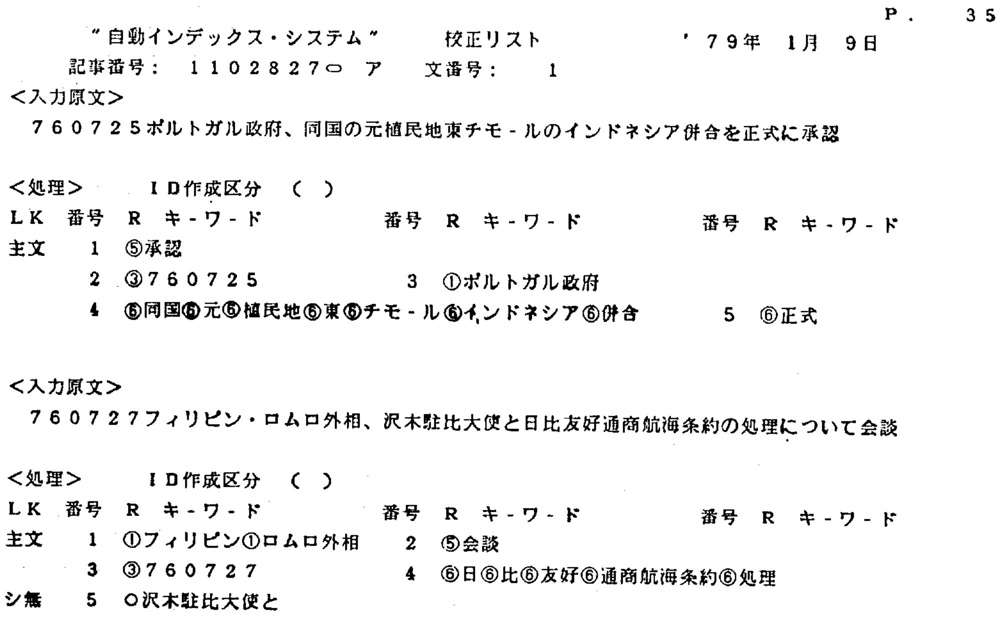
(3) Words in the noun phrase modify the last noun word of the phrase in the analysis.
(4) The "role" is automatically given to each keyword using the results of the above three procedures.

4.3 Man-Machine Interactive Error-Correcting Function

The man-machine interactive error-correction unit consists of a Kanji video terminal and a Kanji line printer.

5. Evaluation of the System

The system has been evaluated by applying it to news information retrieval. The results of this application show, that, based on the assumption that the content word dictionary and the sentence pattern table cover 90% of the processed words and processed sentence patterns, 85 to 90% of the keywords and 80 to 85% of the set roles extracted are estimated to be correct. Also, the time required for indexing is only one third of that required for conventional manual indexing, and the retrieval precision-ratio is improved by 20 to 30% without affecting the recall-ratio. With this method the turn- arround time for information storage is reduced to half of that of the conventional manual method. Examples of output are given in Figure 7.



```
                                                                      P.   35
      "自動インデックス・システム"      校正リスト             '79年 1月 9日
    記事番号: 1102827○ ア   文番号:  1
<入力原文>
  760725ポルトガル政府、同国の元植民地東チモ-ルのインドネシア併合を正式に承認

<処理>      ID作成区分  ( )
LK  番号  R  キ-ワ-ド            番号  R  キ-ワ-ド            番号  R  キ-ワ-ド
主文   1  ⑤承認
       2  ③760725           3  ①ポルトガル政府
       4  ⑥同国⑥元⑥植民地⑥東⑥チモ-ル⑥インドネシア⑥併合      5  ⑥正式

<入力原文>
  760727フィリピン・ロムロ外相、沢木駐比大使と日比友好通商航海条約の処理について会談

<処理>      ID作成区分  ( )
LK  番号  R  キ-ワ-ド            番号  R  キ-ワ-ド            番号  R  キ-ワ-ド
主文   1  ①フィリピン①ロムロ外相   2  ⑤会談
       3  ③760727           4  ⑥日⑥比⑥友好⑥通商航海条約⑥処理
シ無   5  ○沢木駐比大使と
```

Figure 7 Examples of Output

6. Conclusion

A new method of automatic keyword extracting and "role" setting has been proposed and evaluated. An experimental automatic indexing system has been developed utilizing the above mentioned Japanese sentence structure analysis. The analysis is characterized as follows:

(1) It is based on the noun-predicate verb dependency.

(2) Restricted semantic processing becomes possible by utilizing the meaning of a noun as it depends on each predicate verb.

An automatic indexing system has been developed based on the proposed method. By utilizing the system, the following problems which arose with the development of an information retrieval system have been solved, i.e., man-power savings, information storage standardization and the realization of efficient retrieval.

Acknowledgement

The authors wish to thank Professor Toshio Ishiwata of Ibaragi University and Professor Hirohiko Nishimura of Tokyo University of Agriculture and Technology for their helpful discussions. The authors wish to thank Mr. Mieji Shimizu and Mr. Masahiro Sakano of Facom Hitac Ltd. for their encouragement of this study. The authors also wish to thank Dr. Takeo Miura, General Manager of Systems Development Laboratory of Hitachi Ltd., for his farseeing supervision. The authors are also grateful to Mr. Hiraku Hashimoto, Mr. Akira Hakuta and Mr. Kenji Koichi for their computer programming of this study.

References

1. D.G.Hays, "Dependency Theory; A Formalism and Some Observations", Language Vol.40, No.4(1964)
2. C.J.Fillmore, "The Case for Case", Universals in Linguistic Theory, Bach and Harms, eds., Holt, Rinehart, and Winston, New York(1968)
3. Makoto Nagao, Mikio Mizutani and Hiroyuki Ikeda, "An Automatic Method of the Extraction of Important Words from Japanese Scientific Documents", Journal of IPS Japan, Vol.17, No.2(1976 Feb.)
4. Hiroshi Kinukawa, Kenji Tsutsui, Ikuo Odagiri and Mutsuko Kimura, "Stenograph to Japanese Translation System", Information Processing in Japan, Vol.15(1975)
5. Hiroshi Kinukawa and Mutsuko Kimura, "Automatic Indexing System Utilizing Japanese Sentence Analysis", Transactions of IPS Japan, Vol.21, No.3(1980 May)