Linguistic Knowledge Extraction from Real Language Behavior

K.Shirai and T.Hamada

(Department of Electrical Engineering, Waseda University)
(3-4-1 Ohkubo Shinjuku-ku Tokyo Japan)

Abstract -- An approach to extract linguistic knowledge from real language behavior is described. This method depends on the extraction of word relations, patterns of which are obtained by structuring the dependency relations in sentences called Kakari-Uke relation in Japanese. As the first step of this approach, an experiment of a word classification utilizing those patterns was made on the 4178 sentences of real language data. A system was made to analyze dependency structure of sentences utilizing the knowledge base obtained through this word classification and the effectiveness of the knowledge base was evaluated. To develop this approach further, the relation matrix which captures multiple interaction of words is proposed.

1. Introduction

In natural language processing, one of the major problems to be solved is how to describe linguistic and semantic knowledge in the system. If we use no particular technique and capture the behavior in real language as it is, the number of rules, concepts and relations to be arranged may expand so much. But those things contain all essential and primitive elements of language that we want to find out at least. In this paper, it is considered to extract primitive elements from real linguistic behavior, and apply the elements to analysis sentence. As the above-mentioned elements, we use a relation between words. (It is called Kakari-Uke relation in Japanese.)

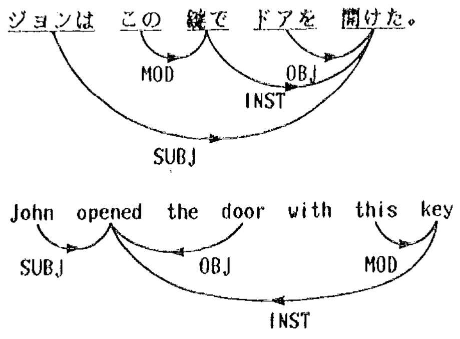

ジョンは この 鍵で ドアを 開けた。
MOD
OBJ
INST
SUBJ
John opened the door with this key
SUBJ
OBJ
MOD
INST

Fig.1 Dependency Relation Structure
(Kakari-Uke Relations)

2. Clustering of Words

2.1. Clustering Method

The process of the word classification based on the pattern of relations is done as follows. First, numbers of sentences are provided and Kakari-Uke relations are given to them. We call those sentences text data. Next we get the source side and the sink side pattern of relations for each word appearing in the text data. Then we calculate a distance between words. The distance is defined as a correspondence between the patterns themselves and the frequency of each relation making the patterns. Words are classified by a clustering algorithm using this distance. The distance has two types; one for the source side patterns and the other for the sink side patterns. For each word, two clustering processes are applied corresponding to those two types of distances. In this paper, the dependency structure is called as the knowledge base.

2.2. Results

We made an experiment of word clustering on the 4178 sentences of text data quoted from computer manuals. In this experiment, a special treatment was taken for compound words to ensure information. There are many compound words in Japanese sentences which are made by combining words and act as one word. They are called Fukugo-go in Japanese. If we treat them all as different from each other, many words appear rarely, so that the relating patterns of each word cannot be captured sufficiently. Because of this reason, we adopted a mechanism that replaces compound words by a normal one including the same meaning grammatical roles in grammar as the former. This mechanism can work automatically as a part of the system.

As the result of this experiment, it was observed as expected that semantically related words tend to be combined. However, some words which have different meaning are combined with a well classified word group, and several well classified groups are combined. Not only synonyms, but also the words similar in some parts as the extension of the words, and also the words which have a common part in the upper concept tend to be combined. It is interesting that antonyms tend to be combined with each other. It was also found that words contained in the same group belong to the same part of speech almost always.

3. Sentence Analysis

3.1. Sentence Analysis System ESSAY

We made ESSAY (Experimental System of Sentence Analysis) which analyzes the dependency structure using the knowledge base. We show the outline of this system in Fig.2. Using the knowledge base, ESSAY analyzes the dependency structure of sentences. If those patterns are used just as they were obtained from the text data, they can only cover the relations which have appeared in the text data. But the clustering process allows the system to cover more relations than appeared in the text data.

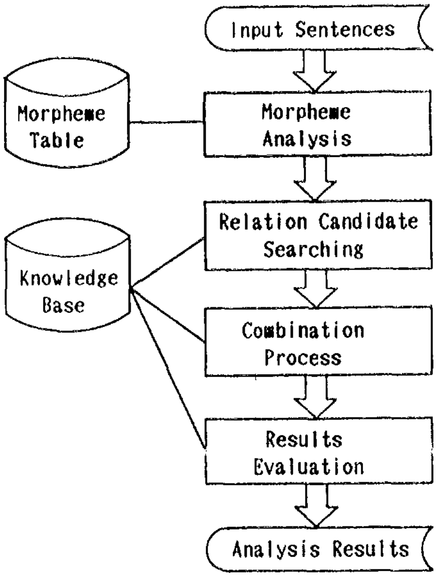


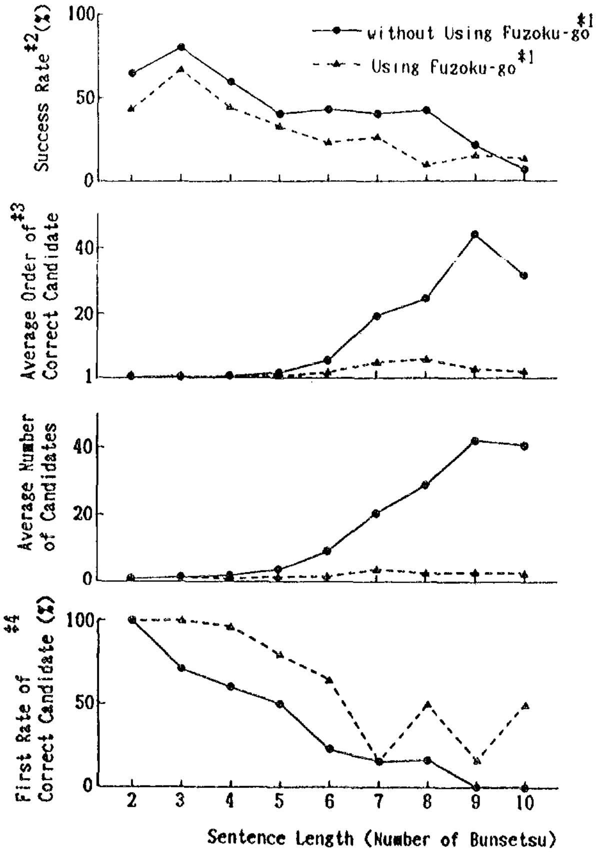
Fig.2 General Flow of ESSAY

```
*****  SENT.NO. =    4 : INPUT IS ...
  図１－３－１－１－４に　ＶＳＡＭカタログの
  機密保護に　関する　パラメタの　関連を　示す
 (The relations of parameters about privacy security
  of VSAM catalogue are shown in Fig.1-3-1-1-4.)

 ****  WORD COMBINATION =    1
  ***  SYNONYM COMBINATION =    1

* EVAL POINT =   90
 ---図１－３－１－１－４に
 !  ( in Fig.1-3-1-1-4 )
 !
 !                ---ＶＳＡＭカタログの
 !                !  ( of VSAM catalogue )
 !                !
 !            ---機密保護に
 !            ! (privacy saecurity)
 !            !
 !        ---関する
 !        !  ( about )
 !        !
 !    ---パラメタの
 !    ! (of parameters)
 !    !
 !--関連を
 ! (the relations)
 !
示す
(are shown)
```

Fig.3 A Sample of Analysis Results

3.2. An Experiment

We made an experiment of sentence analysis with ESSAY. The knowledge base was organized from the 4178 sentences of text data quoted from computer manuals. The input sentences we provided for the test were not contained in the sentences used for knowledge base organization. A sample of the analysis result is shown in Fig.3. There is a possibility that a Bunsetu (a kind of phrase structure element) has several ways of possible division into words and Fuzoku-go. The system tests some combinations of those divisions. In this figure, EVAL POINT indicates the value evaluated for each structure that is calculated from the likelihood of each relations constructing the structure. we can express the conclusion as follows:



#1: The experiment was done under two conditions using and without using Fuzoku-go for analysis in order to examine the effect of Fuzoku-go.
#2: The rate at which the analysis succeeds.
#3: The order of correct candidate in the analysis results.
#4: The rate at which the correct candidate is ranked first.

Fig.4 Analysis Results of every Sentence Length

a) There is a problem that the long sentence with many Bunsetu often makes too many combinations of relation candidates.

b) There are some cases that no result is obtained because only a part of words does not have a relation candidates although all of others have the correct relations.

c) It is difficult to describe a parallel relation using relations between two words. Therefore, it is difficult to analyze a sentence containing parallel relations.

d) The rate at which the analysis succeeds depends on the length of the sentence. As the sentence becomes longer, the rate becomes lower. The average of the rate was about 40 per cent.

This result is shown in Fig.4.

4. More Complicated Data Structure

ESSAY decides the relations according to the connection only between two words. The other parts of the sentence take no role in this decision at all. But the relations complicatedly interact to one another in actual sentences. In this section, we describe how to deal with the interaction of the relations to provide a wider ground for judging propriety of relations.

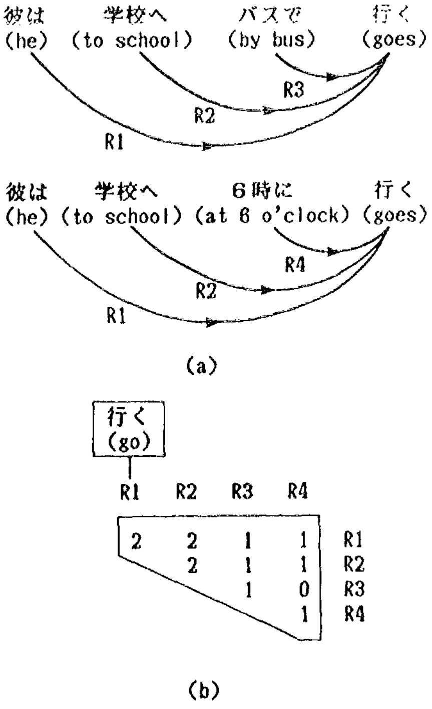


Fig.5 Relation Matrix

4.1. Co-occurrence of Relations

There are words relating to more than two other words at the same time. As shown in Fig.5(a), four kinds of relations appear in the text data. If more than two kinds of relations appear at the same time, the frequency of relations are counted. Then frequency table is expressed by a matrix called relation matrix shown in Fig.5(b). The element Mii means frequency of Ri itself, and the element Mij means frequency of appearance of both Ri and Rj at the same time. This matrix is obtained for each word that have been related with more than two words at the same time. Utilizing this matrix, we can get wider ground for judging propriety of relations. When the relation "go -(to)- school" is obvious, seeing element M2i and Mi2 (i≠2) of the matrix, we can get probability of each relation Ri in this situation.

4.2. Effect of the relation Matrix

Using this matrix, the ground for judging propriety of the relations becomes wider and the number of candidates can be effectively reduced. Secondly, because each relation becomes more reliable, it is expected to get relations according to the sentence meaning.

5. Conclusion

We have introduced a bottom up approach of organization for a linguistic knowledge base. For the organization of knowledge base, continuous human effort has been required. The vocabulary of the knowledge base depends on the quantity of text data.

Linguistic knowledge base organized in this manner may not be so powerful as those constructed analytically. But such method may open an automatic way of the knowledge acquisition and there may be a possibility to discover rules and properties which we have never noticed.

REFERENCE

[1] Bobrow, D.G., and Winograd, T. 1977. An overview of KRL, a knowledge representation language. Cognitive Science 1:3-46.
[2] Woods,W.A. 1973. Progress in natural understanding: An application to lunar geology, AFIPS Conference Proceeding 42, 1973 National Computer Conference. Montvale N.J.: AFIPS Press, 441-450.
[3] Quillian,M.R. 1968. Semantic memory. In Minsky, 227-270.
[4] Fillmore,C. 1968. The case for case. In E.Bach and R.Harms (Eds.), Universals in linguistic theory. New York: Holt, Rinehart, and Winston, 1-88.
[5] Katke,W. 1985. Learning language using a pattern recognition approach. The AI magazine Spring, 1985.
[6] Shirai,K., Hayashi,Y., Hirata,Y., and Kubota,J. 1985. Database formulation and learning procedure for Kakari-Uke dependency analysis.(in Japanese) The transaction of information processing society of Japan, Vol.26, No.4, 706-714.