A Proposal of Paraphrasing Method Using Inductive Learning for Dialogue System

Keigo Kashima, Kenji Araki, *Member, IEEE* and Koji Tochinai

Abstract--**In recent years, many methods of the paraphrasing are proposed since the technology of paraphrasing is able to apply to various applications [1-3] [6-11]. The knowledge based methods for paraphrasing were proposed mainly. However, these methods cannot deal with the situation which is not expected beforehand. And it is difficult to make clear the quantity and the kind of knowledge which should be prepared beforehand. Therefore, we propose the method which generates the paraphrased sentence by using Inductive Learning [13]. This method is robust because the method can adapt to a user dynamically by acquiring paraphrasing rules from examples using Inductive Learning. We adapt this method to the question-answering system [12]. In the question-answering system, it is one of the problems that it cannot answer the questions because they are the different expressions even if they are the same meaning. To solve the problem, the system generates the paraphrased sentences which are the same content and the different expression. By using generated paraphrased sentences in the matching process of the question-answering system, the system based on our proposed method realizes more efficient matching. And it can answer even the question which cannot be answered by the previous methods by using this system. We have constructed the paraphrasing system for the question-answering system on our proposed method, and showed the effectiveness of it in this paper.**

Index Terms— **Inductive Learning, Natural language processing, Paraphrasing, Question-Answering system**

I. Introduction

ORdinary Japanese sentences are expressed by two kinds of characters: i.e. *Kana* and *Kanji*. *Kana* is Japanese phonographic characters and has about fifty kinds. *Kanji* is ideographic Chinese characters and has about several thousand kinds. In natural language, plural expressions for one thing exist. Paraphrasing is work which changes a certain language expression into another expression in the same language preserving its meaning. The method of paraphrasing is not proposed as much as the method of translation which transforms the wording of a sentence or a text into different wording. In recent years, the technology of paraphrasing has been popular. Because the technology of automatic paraphrasing is an important technology which may be applicable to various applications of natural language processing. The knowledge based methods of paraphrasing are proposed. There is the method which uses the knowledge of the Japanese dictionary. And the method which uses the knowledge of syntactic analysis was proposed in research on the past paraphrasing. These are the methods which are preparing the rules and the knowledge for the situation expected beforehand and use them for paraphrasing. However, these methods also have many problems. As first problem, it is sometimes unable to paraphrase the sentence because the rules and the knowledge for paraphrasing are only available for the expected situation. These systems can deal only the situation expected beforehand. As second problem, it is difficult to make clear the quantity and kind of the knowledge and rule which should be prepared beforehand. Thus, there are some problems in the methods of the knowledge based approach. Then, in this paper, we propose the method which generates the paraphrased sentence by using Inductive Learning which acquires rules for paraphrasing recurrently from the actual paraphrased examples. Inductive Learning does not need a huge quantity of a corpus compared with the method of a knowledge based approach and robustness is high. It can adapt to a target dynamically and covers the wide range of paraphrasing. It is difficult to evaluate objectively when the purpose is not set up. Therefore, setting up the purpose of paraphrasing is necessity. Then, we paid attention to dialogue process which performs question-answering. The present question-answering system has many problems. One of them, the system can answer only the question of the expression expected beforehand since the process of the one to one matching. In the process of the one to one matching, the system compares respectively the inputted question sentence with the question sentence which is given to the system beforehand. In this process, a problem occurs. The problem is that it occurs the situation which the system cannot answer since the expression of an inputted sentence is different from the expression expected beforehand, nevertheless the meanings of the question sentence is the same. And it is impossible to prepare the sentence of all expectable expressions for a system beforehand since there is the variation of the expressions. Then, in this paper, we propose using a paraphrased sentence for the question-answering system. The purpose of this system is to learn to answer even the question which cannot be answered by the previous methods by realizing more efficient matching by using generated paraphrased sentences in the matching process of the question-answering system. We performed experiment to evaluation this system.

This work is partially supported by the Grants from the Government subsidy for aiding scientific researches (No.14658097) of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Keigo Kashima and Kenji Araki are with the Graduate School of Engineering, Hokkaido University, Kita 13 Nishi 8, Kita-ku, Sapporo, Hokkaido, 060-8628, Japan. (telephone: +81-11-706-7389, e-mail: {kashima, araki}@media.eng.hokudai.ac.jp)

Koji Tochinai is with the Graduate School of Business Administration, Hokkai Gakuen University, Asahimachi 4-1-40, Toyohira-ku, Sapporo, Hokkaido, 062-8605, Japan. (telephone: +81-11-841-1161, e-mail: tochinai@econ.hokkai-s-u.ac.jp).

The ISBN number for this conference is 1-86467-114-9.

In this paper, we describe the purpose of our research, the results of the performance experiment, and consideration of effectiveness of this system. At last, we describe the conclusion and future problems for it.

II. PROCESSES

A. *Outline*

Fig. 1 shows the process of the system. It consists of five processes that are the answering process, the paraphrasing process, the proofreading process, the learning process and the feedback process. When a question sentence is inputted, the morphological analysis is performed at first by using JUMAN[5] which is the Japanese morphological analysis system. In the answering process, the system find out the question sentence which is the same expression in the sentences registered in the examples of question-answering dictionary on condition that the both of the word strings and the result of morphological analysis of the sentence which is inputted by a user correspond with them of the sentences registered in the dictionary. The pairs of the question sentence and the answering sentence are registered in this dictionary. If it succeeds in finding out the sentence which corresponds with the sentence which is inputted by a user in this process, the system outputs answering sentence according as the examples of question-answering dictionary, and progress to the proofreading process, the learning process and the feedback process. If it fails in finding out the sentence which corresponds with the sentence which is inputted by a user, a user inputs a sentence which is the same semantic content and the different expressions. Here, it acquires a pair of sentences which are the same meanings and the different expressions. The system acquires the paraphrasing rules from the pair of them recurrently. The paraphrasing rule dictionary is updated. The paraphrasing rules are registered in this dictionary. And the paraphrased sentences are generated in the paraphrasing process by using these rules. The system compares the paraphrased sentences with the sentences which are registered in the example of question-answering dictionary. If it succeeds in matching, an answering sentence is outputted according as the examples of question-answering dictionary. As a result, by using generated paraphrased sentences in the matching process of the question-answering system, the system based on our proposed method realizes more efficient matching. And it can answer even the question which cannot be answered by the previous methods.

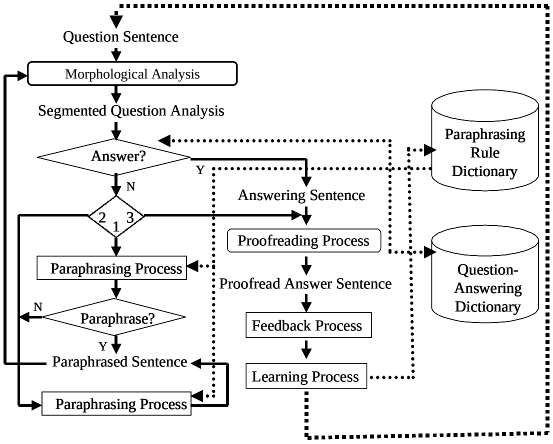

Fig. 1 :The flow of processing

B. *Answering process*

In this process, it compares the question sentence which is inputted by a user and the paraphrased sentences in the paraphrasing process with the sentences registered in the example of question-answering dictionary. If there are the sentences which correspond to the question sentence registered in the example of question-answering dictionary, the answering sentence to the question sentence is outputted according as the dictionary. When both of the word strings and the result of morphological analysis of each sentences correspond completely, these sentences are regarded as the same sentence. It is supposed that matching was successful at this time.

C. *Proofreading process*

In this process, it makes a judgment by a user whether the outputted answering sentence is correct or not to the question sentence which the user inputted. When both of the grammar and the meaning of the answering sentence is correct, the outputted answering sentence is regarded as the correct answering sentence. If the outputted answering sentence is not correct to the question sentence, a user needs to proofread it. Inputted proofread sentence is registered in the examples of question-answering dictionary in the learning process.

D. *Learning process*

In this process, the rules are acquired from the pair of the question sentence which a user inputted. First, a pair of the sentences which a user inputted is acquired as a paraphrasing rule. Next, this rule is compared with the rule which is registered in the paraphrasing rule dictionary and the common part and the different parts are extracted as the common part rule and the different part rule. The common part is defined when the morphemes which include the result of morphological analysis and the word strings are corresponded. The common part is extracted when more than 50% of the number of all the morphemes and the word strings correspond. The common part rule is defined as the rule which the different parts of the rule of cause are transposed to the variables. The part of the rest of the rule is defined as the different part rule. Furthermore, the rules acquired at this time and the rules of the paraphrasing rule dictionary perform the same process, and acquired rules. This process is recurrently performed until a new paraphrasing rule is no longer acquired. A processing procedure which consists of four steps is shown below. And an example of this process is shown in Fig. 2.

Furthermore, the answering sentence which a user inputted in the proofreading process is registered with the example of

*(a): / Sidoni:*n / *made:*p / *no:*p / *koukuuken:*n / *ha:*p / *ikura:*adv / *desu:*p / *ka:*p / *?:*sym /
*(a): / Sidoni:*n / *he:*p / *no:*p / *tiketto:*n / *ha:*p / *ikura:*adv / *suru:*v / *no:*p / *?:*sym /
(How much is the ticket to Sydney?)
*(b): / LA:*n / *made:*p / *no:*p / *koukuuken:*n / *ha:*p / *dono:*d / *kurai:*adj / *?:*sym /
*(b): / Losanzerusu:*n / *he:*p / *no:*p / *tiketto:*n / *ha:*p / *dono:*d / *gurai:*adv / *?:*sym /
(How much is the ticket to Los Angels?)

<Common parts Rules>
*(c): / @0 / made:*p / *no:*p / *koukuuken:*n / *ha:*p / *@1 / : / @0 / he:*p / *no:*p / *tiketto:*n / *ha:*p / *@1 /*

<Different parts Rules>
*(d): / Sidoni:*n / *: / Sidoni:*n /
*(e): / LA:*n / *: / Losangerusu:*n /
*(f): / ikura:*adv / *desu:*p / *ka:*p / *?:*sym / *: / ikura:*adv / *suru:*v / *no:p / ?:*sym /
*(g): / dono:*d / *kurai:*adj / *?:*sym / *: / dono:*d / *gurai:*adv / *?:*sym /

n:Noun, p:Particle, adj:Adjective, adv:Adverb, v:Verb, d:Determiner, sym:Symbol, @:Variable

Fig. 2 : Example of rule acquisition

*(a): / kujira:*n / *ha:*p */sakana:*n / *desu:*dec / *ka:*p / *?:*sym /
*(a): / kujira:*n / *ha:*p / *sakana:*n / *desyou:*dec / *ka:*p / *?:*sym /
(Is a whale a fish?)

<End of sentence Rule>
*(b): / @0 / desu:*dec / *ka:*p / *?:*sym /
*(b): / @0 / ka:*p / *?:*sym /

n:Noun, p:Particle, dec:Decision word, sym:Symbol, @:Variable

Fig. 3 : Example of the end of the sentence rule

question-answering dictionary, and this dictionary is updated.

① Select
The pair of a rule which corresponds with 50% or more is selected. When more than 50% of the number of all the morphemes which include the result of morphological analysis and the word string correspond, the pair is selected.

② Generation
The part which correspond is extracted as the common part rule by adding the different part transposed to a variable in the selected rule. And the different parts are extracted and they are defined as the different part rule.

③ The new paraphrasing rules which are generated are registered in the paraphrasing rule dictionary and process of (①,②) is repeated.

④ It finishes when a new rule is not generated.

Moreover, the end of the sentence rule is extracted. First, the independent words which are at the end of sentences which are inputted by a user are extracted. Next, if these two independent words which are extracted correspond, the end of the sentence rule is extracted. The end of the sentence rule is defined as the rule of the part at the back from the independent word. Fig. 3 shows an example of acquisition of the end of the sentence rule.

Fig. 2 shows an example of the acquisition of the rules. In this process, our system extracts a common part rule and a different part rule according to word strings and the result of morphological analysis. " */ @0 / made:*p */ no:*p */ koukuuken:*n */ ha:*p */ @1 /* " and " */ @0 / he:*p */ no:*p */ tiketto:*n */ ha:*p */ @1 /* " ,example(c) are extracted from example(a) and example(b) as common part rules. The mark "*@0* and *@1* " means a variable. And example(d) " */ Sidoni:*n */* " and " */ Sidoni:*n */* ", example(e) " */ LA:*n */* " and " / *Losangerusu:*n */* ", example(f) " */ ikura:*adv */ desu:*p */ ka:*p */ ?:*sym / " and " / *ikura:*adv */ suru:*v */ no:p / ?:*sym / " and example(g) " */ dono:*d */ kurai:*adj */ ?:*sym / " and " */ dono:*d */ gurai:*adv */ ?:*sym / " are extracted from example(a) and example(b) as different part rules.

<Input sentence>
(a): / anata:n / ha:p / itsu:adv / hirugohan:n / wo:p / tabe:v / masu:suf / ka:p / ?:sym /
(What time do you eat the lunch?)

<Rules>
(b): / @0 / ha:p / itsu:adv / hirugohan:n / wo:p / @1 / : / @0 / ha:p / nanzi:n / ni:p / hirugohan:n / wo:p / @1 /
(c): / kimi:n / : / anata:n /
(d): / tabe:v / masu:suf / ka:p / ?:sym / : / tori:v / masu:suf : / ka:p / ?:sym /

<Generation of paraphrased sentence>
(e) / @0 / ha:p / itsu:adv / hirugohan:n / wo:p / @1 / : / @0 / ha:p / nanzi:n / ni:p / hirugohan:n / wo:p / @1 /
(f): / anata:n / ha:p / itsu:adv / hirugohan:n / wo:p / @1 / : / kimi:n / ha:p / nanzi:n / ni:p / hirugohan:n / wo:p / @1 /
(g): / anata:n / ha:p / itsu:adv / hirugohan:n / wo:p / tabe:v / masu:suf / ka:p / ?:sym /
(g): / kimi:n / ha:p / nanzi:n / ni:p / hirugohan:n / wo:p / tori:v / masu:suf : / ka:p / ?:sym /
(What time do you eat the lunch?)

n:Noun, p:Particle, adj:Adjective, adv:Adverb, v:Verb, suf:Suffix, sym:Symbol, @:Variable

Fig. 4 : Example of generation

Fig. 3 shows an example of the acquisition of the end of the sentence rule. First, the independent words which are at the end of the sentences which are inputted by a user are extracted. In this example, " / sakana:n / "and " /sakana:n / " are extracted each other. These two independent words correspond. And, the part at the back from these words, " / @0 / desu:dec / ka:p / ?:sym / " and " / @0 / ka:p / ?:sym / " are extracted as the end of the sentence rule as the example(b).

E. Paraphrasing Process

In this process, the paraphrased sentence which is the same meaning and the different expression is generated by using the rules which are registered in the paraphrasing rule dictionary. First, the rules which include the variables are extracted from the paraphrasing rule dictionary. The rule in this dictionary consists of a pair of sentence. The rules which is registered in the paraphrasing rule dictionary put into the part of variable in the rule which is extracted. This process is repeated until the part of variable is put. When one of a pair of the rule becomes a question sentence which is inputted by a user, a paraphrased sentence is generated in another of the rule. By applying the process to all the rules which are registered in the paraphrasing rule dictionary, the sentence which is inputted by a user is paraphrased. Fig. 4 shows an example of generation of a paraphrasing sentence.

Fig. 4 shows an example of the generation of a paraphrased sentence. Example(a) " / anata:n / ha:p / itsu:adv / hirugohan:n / wo:p / tabe:v / masu:suf / ka:p / ?:sym / " is inputted sentence. In this figure, there are the acquired rules (b)~(d) in the paraphrasing rule dictionary. First, rule(c) is substituted for "@0" in the rule(b) and the rule(f) is generated. Next, rule(d) is substituted for "@1" in the rule(f). As the result, rule(g) " / anata:n / ha:p / itsu:adv / hirugohan:n / wo:p / tabe:v / masu:suf / ka:p / ?:sym / " and " / kimi:n / ha:p / nanzi:n / ni:p / hirugohan:n / wo:p / tori:v / masu:suf : / ka:p / ?:sym / " are generated. The rule has no variable and one side of the rule " / anata:n / ha:p / itsu:adv / hirugohan:n / wo:p / tabe:v / masu:suf / ka:p / ?:sym / " is equal to the inputted sentence. And the another side of the rule " / kimi:n / ha:p / nanzi:n / ni:p / hirugohan:n / wo:p / tori:v / masu:suf : / ka:p / ?:sym / " is generated paraphrased sentence to the inputted sentence.

III. Experiment and Evaluation

A. Data

The first author prepared 200 sentences from *YOTEN DON*[4] which is an English reference book for junior high school and answering sentences. And these sentences are registered in the example of question-answering dictionary as an initial state of this dictionary. Question sentences are prepared as the sentence which is inputted at the experiment. These question sentences are the same meanings and the different expressions as the sentences which is registered in the example of question-answering dictionary at first. These question sentences is used for inputted sentences.

B. Procedure

The initial state of the paraphrasing rule dictionary empty. In Inductive Learning, there is no necessity of preparing knowledge in advance and the rules can be acquired as the learning dynamically. When the user gives the system a question sentence, the system attempts to generate the paraphrased sentence using the paraphrasing dictionary. After the paraphrased sentences are generated, the answering sentence is

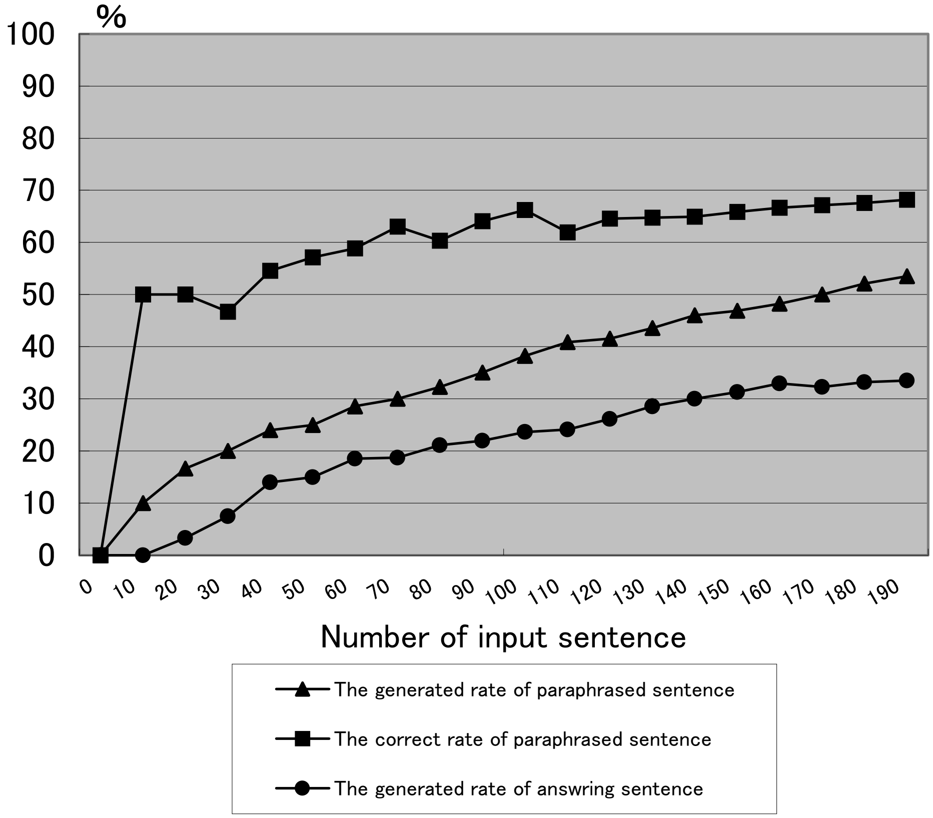

Fig. 5 : Transition of evaluation

outputted using the examples of question-answering dictionary.

C. *Standards for Evaluation*

In these experiment, we define the evaluation formulas to evaluate the experiment result. The generated rate of the paraphrased sentence, the correct rate of paraphrased sentence, the generated rate of the answering sentence and the correct rate of answering sentence are defined as follows:

$$\textit{Correct rate of generated sentence [\%]} = \frac{\textit{Number of correct paraphrased sentence}}{\textit{Total of generated paraphrased sentence}} \times 100$$

$$\textit{Generated rate of answering sentence [\%]} = \frac{\textit{Number of outputted answering sentence}}{\textit{Total of inputted sentence}} \times 100$$

$$\textit{Correct rate of answering sentence [\%]} = \frac{\textit{Number of correct answering sentence}}{\textit{Total of generated answering sentence}} \times 100$$

$$\textit{Generated rate of paraphrased sentence [\%]} = \frac{\textit{Number of generated paraphrased sentence}}{\textit{Total of inputted sentence}} \times 100$$

We have carried out the performance evaluation with four rates. In this experiment, a correct paraphrased sentence is defined as the sentence whose grammar is correct and a correct answering sentence is defined as the sentence whose grammar is correct and the meaning of the sentence is suitable for the question sentence. The generated rate of the paraphrasing sentences is defined as the rate of the number of generated paraphrasing sentences in the number of the inputted sentences. The corrected rate of the paraphrasing sentences is defined the rate of the number of correct paraphrasing sentences in the number of the generated paraphrasing sentences. The outputted rate is defined the rate of the number of outputted response answering sentences in the number of the inputted sentences. The corrected rate of the outputted answering sentence is defined the rate of the number of correct outputted response answering sentences in the number of the outputted response answering sentences to the inputted question sentences. Evaluation of this system is performed these four rates.

D. Results

Table 1 shows the result of the experiment. Fig. 5 shows the transition of evaluation. The generated rate of the paraphrased sentence is 57.50% and the correct rate of the paraphrased sentence is 68.18%. The correct rate of paraphrased sentence increases as the input data increases. And the average of the number of the paraphrasing sentence were 2.41 sentences. It shows that the system has acquired the rules of paraphrasing. The system has improved as the learning advances. And, the generated rate of the answering sentence is 33.50% and the correct rate of the answering sentence is 95.74%.

Table 1 : Result of the experiment

	Number	%
Input	200	–
Paraphrased sentence	115	57.50
Correct paraphrased sentence	79	68.18
Answering sentence	67	33.50
Correct answering sentence	64	95.74

IV. Consideration

As a result of this experiment, the generated rate of the paraphrased sentence is 57.50% and the correct rate of the paraphrased sentence is 68.18%. In the dialogue process of question-answering, the system can answer even the question which cannot be answered by the previous methods since the system generate the paraphrased sentence. And the effectiveness of this system was shown. The generated rate of the paraphrasing sentence increased as the learning advances. The reasons for this are that the number of the paraphrasing rules are increasing as the learning advances. Moreover, although the correct rate of paraphrased sentence was low in the beginning, the correct rate of it increased. This is considered because the feedback process is performed correctly. As the result, the correct paraphrasing rules are applied and the incorrect paraphrasing rules are no longer applied. The correct rate of answering sentence is high as 95.74%. The reason for this is that the answering sentence to the question sentence is outputted only when both of the word strings and the result of morphological analysis of the question sentence which is paraphrased correspond with the question sentence registered in the example of question-answering dictionary exactly. And the question sentence which is the different meaning from the question sentence which is inputted by a user is hardly generated by applying the incorrect rules. As the result, the correct rate of the answering sentence was very high and the generated rate of the answering sentence was not so high.

V. Conclusion

In this paper, we described that the technology of paraphrasing is important in the natural language process and the technology can apply the various applications. Then, we have described the method which the paraphrasing sentences that are generated by Inductive Learning and adapts to the dialogue process of question-answering. The system realizes more efficient matching and can answer even the question which cannot be answered by the previous methods. One of the problems of question-answering system is that the system cannot answer the question which is the different expression even if it is the same meaning. Our system is the system to solve this problem. The purpose of this system is to learn to answer even the question which cannot be answered by the previous methods by realizing more efficient matching by using generated paraphrased sentences in the matching process of the question-answering system. We carried out performance evaluation experiment. The paraphrased sentences are generated and the system can answer the question which cannot be answered.

For the future works, we will consider in learning process whether there is some information which is available in addition to the present information to get the paraphrased rules on the common part and the different part. In addition, more detailed evaluation is necessary.

References

[1] Masaki Murata and Hitoshi Isahara "Universal Model for Paraphrasing-Using Transformation Based on a Defined Criteria", *In Proceedings of NLPR2001 Workshop*, pp. 47-54, 2001.

[2] Regina Barzilay and Kathleen R. McKeown, "Exacting paraphrases from a parallel corpus", *In Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pp. 50-57, 2001.

[3] Tetsuro Takahashi, Tomoya Iwakura, Ryu Iida, Atsushi Fujita and Kentaro Inui "KURA: A Transfer-Based Lexico-Structural Paraphrasing Engine", *In Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pp. 37-46, 2001.

[4] *Gakken, "YOTEN-DON Tyugaku eigo", Gakusyu Kenkyusya*, 2001

[5] Sadao Kurohashi and Makoto Nagao, *Japanese Morphological Analysis System JUMAN version3.5*. Department of Informatics, Kyoto University. (in Japanese)

[6] Noriko Tomuro and Steven L. Lytinen "Selecting Features for Paraphrasing Question Sentences", *In Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pp. 55-62, 2001.

[7] Y.Matsuo, S.Shirai and S/Ikehara "Changing Syntactic Classes is Transfer-based Machine Translation. ", *In Proceedings of the 3rd Natural Language Processing Pacific Rim Symposium, pp.432-437, 1995*

[8] P.Brown, S. Della Pietra, V. Della Pietra, and R. Mercer "The mathematics of statistical machine translation: Parameter estimation", *Computational Linguistics, 19(2), pp. 263-311.*

[9] Kentaro Torisawa "A Nearly Unsupervised Learning Method for Automatic Paraphrasing of Japanese Noun Phrases", *In Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pp. 63-71, 2001.

[10] Agichtei, E., S. Lawrence, and L. Gravano "Learning search engine specific query transformations for question answering", *In Proceedings of 10th International World Wide Web Conference(WWW10)*, Hong Kong, 2001

[11] Weizenbaum, J. : ELIZA – A Computer Program for the Study of Natural Language Communication Between Man And Machine, *Communications of the Association for Computing Machinery*, vol.9, No.1, pp. 36-45, 1966.

[12] K.Araki and K.Tochinai："Effectiveness of Natural Language Processing Method Using Inductive Learning", *Proceedings of the IASTED International Conference ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING*, pp.295-300, 2001.