Information Processing and Management 40 (2004) 957–972

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Word classification and hierarchy using co-occurrence word information

Kazuhiro Morita, El-Sayed Atlam [*], Masao Fuketra, Kazuhiko Tsuda, Masaki Oono, Jun-ichi Aoe

Department of Information Science and Intelligent Systems, University of Tokushima, Tokushima 770-8506, Japan

Received 24 March 2003; accepted 25 August 2003
Available online 8 October 2003

Abstract

By the development of the computer in recent years, calculating a complex advanced processing at high speed has become possible. Moreover, a lot of linguistic knowledge is used in the natural language processing (NLP) system for improving the system. Therefore, the necessity of co-occurrence word information in the natural language processing system increases further and various researches using co-occurrence word information are done. Moreover, in the natural language processing, dictionary is necessary and indispensable because the ability of the entire system is controlled by the amount and the quality of the dictionary. In this paper, the importance of co-occurrence word information in the natural language processing system was described. The classification technique of the co-occurrence word (*receiving word*) and the *co-occurrence frequency* was described and the classified group was expressed hierarchically. Moreover, this paper proposes a technique for an automatic construction system and a complete thesaurus. Experimental test operation of this system and effectiveness of the proposal technique is verified.

Keywords: Natural language processing; Co-occurrence word information; Co-occurrence frequency; Hierarchically

1. Introduction

How does the human begin accumulate the knowledge and what does he think, when people read or understood *natural language* (NL) documents. To replace these meaning of words with

[*] Corresponding author. Current address: Department of Information Science and Intelligent Systems, University of Tokushima 770-8506, Japan, permanent address: Department of Statistics and Computer Science, Faculty of Science, Tanta University, Tanta, Egypt.
E-mail addresses: atlam@is.tokushima-u.ac.jp, satlam@yahoo.com (E.-S. Atlam).

0306-4573/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2003.08.009

the machine processing it is necessary to analyze the deep meaning of word that exists in the background. However, a huge and a variety information are needed for that, so it is difficult to replace these meaning with the machine processing easily. The thesaurus (control language dictionary) that is a typical knowledge representation solves this problem. In this thesaurus the word is classified and hierarchy by the meaning. So, thesaurus is important for a large amount document classified by the semantic content, but a hand-made thesauruses often have problems such as costs the labor and reflected the subjectivity caused by the manufacturer.

Recently various methods for automatically construction thesaurus (hierarchically clustering words) based on document data on the Web have been proposed (Hindle, 1990; Brwn et al., 1992; Petira, Tishby, & Lee, 1993; Tokunaga, Iwayama, & Tanaka, 1995). The realization of such an automatic construction method would make it possible to (a) save the cost of constructing a thesaurus by hand, (b) do away with the subjectivity inherent in a hand-made thesaurus (c) make it easier to adapt a *natural language processing* (NLP) system to a new domain. The lexical knowledge extracted by co-occurrence information (data) that used as important information to cancel the ambiguous syntax and polysemy. It is also used for clustering the related word meaning together.

Yokoyama and Shinichiro (1998) have proposed a method for classifying the meaning of verbs using co-occurrence word information and some of these verbs can not be classified correctly even an excellent result is comparatively obtained. Moreover, (Li & Abe, 1997) solved the problem of automatically clustering words by estimating a joint distribution over the Cartesian product of a partition of set of nouns (in general, any set of words) and a partition of set of verbs (e.g., noun set and verb set) and proposed an estimation algorithm based on *minimum description length* (MDL) principle. The MDL principle is a well-motivated and theoretically sound principle data compression that estimated from information theory and statistics. As a strategy of statistical estimation, MDL is guaranteed to be near optimal. Moreover, the efficiency processing of the description length is devised and alternately merging both word sets to make the thesaurus. However, the system made by Li and Abe (1997) still becomes practical use as only a sight evaluation is done and making only classification for words.

In this research, an automatic thesaurus is classified and hierarchy knowledge that human begin has constructed using the co-occurrence word (*receiving word*) and the co-occurrence frequency. When a word meaning distance is obtained from the co-occurrence relation, the word that co-occurs mutually with the same relation as a certain *receiving word* is assumed to be similar. Under this assumption, one group can bring the word group with the common feature by taking the similarity between them together. Moreover, the super-concept and sub-concept relation are given between groups by taking the similarity between each group. Therefore, when this system is characterized, the word is not only classified, but also making the hierarchy of them.

Section 2 explains *co-occurrence word information* in detail. Section 3 explains the outline of the system that becomes basic of this research. In section 4, the system is evaluated by the experimental results. Section 5 describes conclusion and possible future work.

2. Co-occurrence word information and storage technique

This section confirms the *co-occurrence word information* meanings by describing the outline about these *co-occurrence word information* and the dictionary that becomes basic of the NLP.

2.1. Co-occurrence word information

The utility of *co-occurrence word information* in the NLP system is extremely high. It is very important for canceling the ambiguous and the polysemy of words to improve the accuracy of the entire system. Various researches (Fukumoto & Tsuji, 1994; Kobayashi, Tokunaga, & Tanaka, 1996; Takahashi & Itabashi, 1998) are done for that using co-occurrence word information. Moreover, people have collected *co-occurrence word information* and the difference is seen by collection person's aspect. Our aim is to improve the accuracy of *co-occurrence information* though there is a limit amount of collection with a time restriction. This *co-occurrence word information* extracts from a large amount of corpus by the objective method and there are a lot of researches (Matsumoto, 1992; Morimoto, Iriguchi, & Aoe, 1993; Morita, Mochizuki, Yoshihiro, & Aoe, 1998; Yokoyama & Shinichiro, 1998) that use these results. Moreover, the information that defines meaning relation usually exists between two words is as follows:

Definition 2.1. When the related information α is defined between two basic words X and Y.

It is written: (X, Y, α)

The related information α has a variety of definition and retrieval demand as follows:

(a) *Relation between super-concept and sub-concept hierarchy* (*hierarchical relationship*)

The classification (concept hierarchy) represented by the thesaurus is a very simple knowledge representation and a very wide range of the application. A basic inference of this expression is because of the super-concept (high rank position) and the sub-concept becoming *co-occurrence word information* such as the concept "Clothes" and "Sports shirt". Moreover, ("Country Name" and "America"), ("Country Name" and "Canada") can also possible to define as *co-occurrence word information.*

(b) *Relation between verb and noun phrase in the case structure* (*case relation* [Hirao & Matsumoto, 1994])

In the case structure's storing meaning restriction of the noun phrase to the verb is obtained. For instance, case relation (run, dog, subject) can be obtained from (dog, animal, super-concept) and (run, animal, subject). Therefore, case relation (run, animal, subject) and (run, car, subject) are obtain.

(c) *Compound word relation*

The compound word "Canadian Nationality" is invented as ("Country Name" and "Nationality") also ("America" and "United States of America") is a compound word.

(d) *Synonym relation*

"America" and "United States of America" are synonym, also, "Cutter" and "Sports shirt" are shortening word of synonym.

Although, the surface case likes the nominative case, the objective, and the possessive, etc. can be easily decided from a large amount of corpus at the present stage. The deep case that thinks what role other words have for the verb cannot be extracted as related information. Therefore, an enough analysis cannot be done in this present structural analysis case. Moreover, Definition 2.1 is enhanced from the consideration of the frequency data of the co-occurrence (co-occurrence frequency) which define as follows:

Definition 2.2. Related information α is defined between two words X, Y, and *co-occurrence frequency* f as follows:

$$(X, Y, \alpha, f)$$

Example 2.1. The relation (‘‘Chirp’’, ‘‘Bird’’,‘‘2’’) is the relation between the subject and the predicate (‘‘Bird’’ and ‘‘Chirp’’) with *co-occurrence frequency* 2.

In this example, the word ‘‘Bird’’ is called a *lying word* and the word ‘‘Chirp’’ is called a *receiving word*. Moreover, related information (particle) is called co-occurrence relation labels.

2.2. Dictionary constructing method

The linguistic knowledge is very important in NLP system and does not limit to co-occurrence information. To make the computer analyze and generate the language, it is necessary to save the linguistic knowledge, therefore computer can be able to use it.

In NLP when analyzing the sentence and generating it, knowledge concerning the language must be brought together as a grammatical rule collection dictionary. If the word that not found in the dictionary is used in NLP system, the information in the dictionary is contradicted (i.e., an unabashedly wrong interpretation is done) and the judgment is stopped. Therefore, taking out the linguistic knowledge of the dictionary system becomes the most basic operation needed at all stages of NLP. The ability of NLP is controlled greatly by the amount of the dictionary composition and the quality.

Because the linguistic knowledge has a various classifications method, so dictionary is made at each classification. Also, the NLP system selects the plural and uses a necessary dictionary. Moreover, the collocation dictionary that brings *co-occurrence word information* together is roughly importance now. A lot of researches (Koyama & Aoe, 1995) on the dictionary constructing method and the dictionary system are done.

3. Classification and hierarchy of word

3.1. System overview

Fig. 1 shows the system overview chart in this research. This system divides roughly into three processing: (a) co-occurrence information registration (b) the word information making and (c) hierarchy.

Example 3.1. Sets of co-occurrence information are

$$\begin{aligned}C1 = \{&(\text{“}\textit{Swim}\text{”}, \text{“}\textit{Young person}\text{”}, \text{“2”}), (\text{“}\textit{Run}\text{”}, \text{“}\textit{Young person}\text{”}, \text{“7”}),\\ &(\text{“}\textit{Shout}\text{”}, \text{“}\textit{Young person}\text{”}, \text{“1”}), (\text{“}\textit{Speak}\text{”}, \text{“}\textit{Young person}\text{”}, \text{“5”}),\\ &(\text{“}\textit{Swim}\text{”}, \text{“}\textit{Elderly person}\text{”}, \text{“3”}), (\text{“}\textit{Run}\text{”}, \text{“}\textit{Elderly person}\text{”}, \text{“9”}),\\ &(\text{“}\textit{Speak}\text{”}, \text{“}\textit{Elderly person}\text{”}, \text{“2”}), (\text{“}\textit{Shout}\text{”}, \text{“}\textit{Dog}\text{”}, \text{“6”}),\\ &(\text{“}\textit{Speak}\text{”}, \text{“}\textit{Dog}\text{”}, \text{“8”}), (\text{“}\textit{Run}\text{”}, \text{“}\textit{Car}\text{”}, \text{“15”})\}\end{aligned}$$

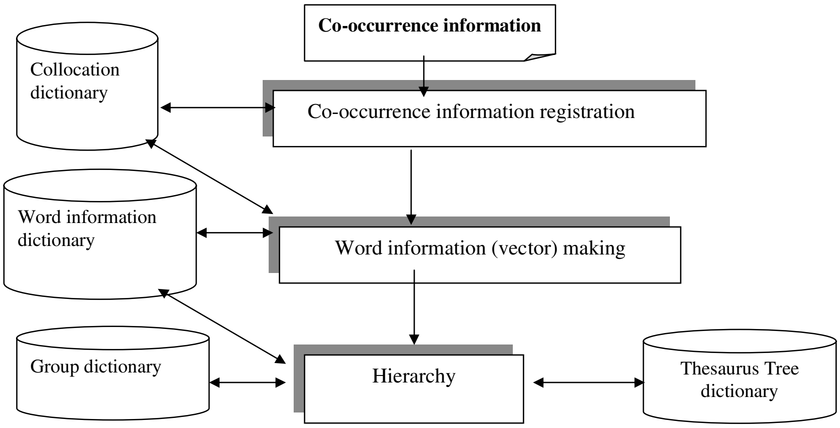
Co-occurrence information
Collocation dictionary
Co-occurrence information registration
Word information dictionary
Word information (vector) making
Group dictionary
Hierarchy
Thesaurus Tree dictionary

Fig. 1. System overview chart.

3.1.1. *Co-occurrence information registration*

In the co-occurrence information registration: *receiving word* and *lying word* are registered as co-occurrence information.

When you register co-occurrence information as in the Example 3.1 in the link trie, the distinction between the *receiving word* and *lying word* does not attach because two words are registered. Then, identifier (object-marker) * of the *lying word* is inserted in the head of the key. For instance, the co-occurrence information in Example 3.1 ("Swim", "Young person", "2") are registered by "**Young person*" and "Swim" as keys.

3.1.2. *Word information making*

Word information is made collectively by comparing information that shows the feature of the word mutually in some shape. In this research, to store word information the inverted index method and the vector method are expressed as a data method. As the reason, it is given to treat the vector method by the set theory, so it is easy in the programming to take AND and OR operator. Moreover, time and processing that hangs comparing will be reduced because it can have word information by the bit row. Thereafter, word information on the vector form which is composed by *receiving word* and *co-occurrence frequency* is called *Word vector*.

3.1.3. *Hierarchy*

It easily explains the flow of processing in *hierarchy*. First of all, the vector which is made by the *word information* is investigated to which node is similar. As a search procedure, the vector group which is a vector for the object node is made and the similarity measurement for Word vector is measured. When the tree search is finished it contains node N that have highest similarity measurement or an addition sub-concept node (called the child node) to the node N. The addition leads to the construction of the tree and the high similarity retaining becomes possible between the super-concept and the sub-concept nodes.

3.1.4. Dictionary composition

The system in this research is composed by four system dictionaries: *collocation dictionary*, *word information dictionary*, *group dictionary* and *thesaurus tree dictionary*.

(1) *Collocation dictionary* is the management of the co-occurrence word information by link tri structure,
(2) *Word information dictionary* (temporary dictionary) is the management of the Word vector. When all processing ends by the system this dictionary deleted,
(3) *Group dictionary* is the management of the Group vector,
(4) *Thesaurus tree dictionary* is the management of relationships between nodes and leaf node in the tree.

3.2. Vector information

3.2.1. Vector word information

Word information vector form is made according to the *receiving word* and the *co-occurrence frequency*. Word information vector W_VEC of a word with the *receiving word* of t times appearance in document can be shown as follows:

$$W\_VEC = \sum_{i=1}^{t} p_i V_i \tag{1}$$

where p_i is *co-occurrence frequency*, and V_i is a vector corresponding to the *receiving word.*

All elements of the vector V_i are symbol string of 0 or 1. To add the vector of a new receiving word is just to adjust one value of symbol string that are changed from 0 to 1 newly, at the same time a linearly independent vector is added newly to the current vector.

Word information dictionary is not an important dictionary for this system because the Word vector can be made at any time based on the co-occurrence relation. However, when doing classification and clustering, this dictionary is assumed as a temporary dictionary and its information becomes important. Therefore, the *word information dictionary* is deleted with all processing terminations.

3.2.2. Group vector

The *Group* vector shows the feature of the word group and this *Group* vector is made from the *Word* vector of two or more words that belong to each group. Therefore, the word group of each group can be the group of words with same feature. Moreover, from the assumption of the Word vector, the *Group* vector is same as the Word vector when the basic group is contained only one word.

The *Group* vector information G of the *receiving word* of s times is shown as

$$G\_VEC = \sum_{i=1}^{s} q_i V_i \tag{2}$$

where q_i the sum of the total *co-occurrence frequency*, and V_i is a vector corresponding to the *receiving word.*

The difference between expression (2) and expression (1) is in the point that two or more words exist in the group. It becomes the co-occurrence relation to all words with compositions *co-occurrence frequency* of them.

3.3. Similarity measurement

The word is classified by using *Word* vector *W_VEC*. The Word vector and the similarity measurement with the vector group to each node are calculated. The cosine angle of two vectors is used in this research though the similarity measurement of two vectors can be defined in various shape in vector space as

$$x \cdot y = |x||y| \cos \alpha \tag{3}$$

where $|x|$ is the length of vector and α is the angle that this vector does. When such a similarity measurement is adopted (Atlam, Fuketa, Morita, & Aoe, 2000; Atlam, Fuketa, Morita, & Aoe, in press; Fuketa, Lee, Tsuji, Okada, & Aoe, 2000; Zhang & Rasmussen, 2001). The similarity measurement of *Group* vector *G_VEC* of *Word* vector *W_VEC* becomes:

$$\mathrm{sim}(W\_VEC, G\_VEC) = W\_VEC \cdot G\_VEC = \sum_{i,j=1}^{t} p_i q_j V_i \cdot V_j \tag{4}$$

Vector V corresponding to the key word of t times, it is assumed to be orthogonal respectively:

$$V_i \cdot V_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

Therefore, $\mathrm{sim}(W\_VEC, G\_VEC)$ is simplified like the following expression:

$$\mathrm{sim}(W\_VEC, G\_VEC) = \sum_{i=1}^{t} p_i q_i \tag{5}$$

However, because inner product is influenced by the magnitude of a vector, the similarity measurement based on this inner product normalized by using the cosine angle of these two vectors and the normalized similarity measurement $\mathrm{sim}'(W\_VEC, G\_VEC)$ will be

$$\mathrm{sim}'(W\_VEC, G\_VEC) = \frac{\mathrm{sim}(W\_VEC, G\_VEC)}{|W\_VEC||G\_VEC|} \tag{6}$$

The similarity measurement in which word information and node information were compared using vector is defined above. Moreover, similarity measurement uses the cosine that *Word* vector *W_VEC* and *Group* vector *G_VEC* of the group are similar.

Example 3.2. *Word* vector to word "Elderly person" of Example 3.1 is

$$W\_VEC = (3, 9, 0, 2)$$

Group vector composed only of word "Young person" is

$$G\_VEC = (2, 7, 1, 5)$$

The similarity measurement of *W_VEC* and *G_VEC* by expression (6) is

$$\mathrm{sim}'(W\_VEC, G\_VEC) = \frac{2*3+7*9+5*2+1*0}{\sqrt{2^2+7^2+5^2+1^2}*\sqrt{3^2+9^2+2^2}} \cong 0.92$$

4. Experiments and evaluation

In this section the method of automating construction thesaurus, the preliminary experiment for testing program and the effectiveness of this system is proposed. Moreover, in the actual experiment small-scale knowledge is initially constructed in and the system verified whether an accurate classification could be done.

4.1. Preparation

The system is constructing with addition of *co-occurrence information* on a constant amount two or more times. The number of system execution to make the classification tree correctly is important and depending on *co-occurrence information*. By some judgments it is necessary to terminate the execution automatically. To solve this problem the *stability of system* will be used. Also, it is possible that a word moves to various groups. However, if the feature of the group is specific and a certain word classified into this group then this word does not move to other groups (i.e., it is stagnated to this group). Therefore, if all words are stagnating the classification tree will be steady. So, the stagnation level that expressed the degree of the stagnation numerically is introduced as follows:

Definition 4.1

$$\text{Stagnation level} = \frac{\text{The number of word which their nodes not move}}{\text{The total number of target words classified by system}} \times 100[\%]$$

The stagnation level means the ratio in which the node is not moved even if the change in a new *receiving word* and frequency occur to the word (*lying word*). Theoretically, the stagnation level increases when the number of system execution increases.

4.2. Preliminary experiment

4.2.1. Observed data

Table 1 shows the data used by the preliminary experiment. In this experiment "feature number" is a number of *the receiving word* and the sign "<>" means the word concept.

In Table 1(a, b, c) *co-occurrence information* related to "Vehicle", "Facilities" and "Musical Instruments" is used. The co-occurrence relation of sub-concepts "Land Vehicle" and "Ship" of the super-concept "Vehicle" is prepared which is used as a *lying word*. Moreover, the words exist in these sub-concepts are used also as *co-occurrence information* for the evaluation. All these

Table 1
Observed data

Concept name	Feature number	Number of words exist in concepts	Example of word
(a) Co-occurrence information related to "Vehicle"			
<Vehicle>	217	35	Vehicle, international flight, and round trip mail
<Land Vehicle>	737	187	Curves, and private car
<Four Wheels>	22	49	Light tiger, benz, and wagon
<Two Wheels>	6	12	Motorcycle and motor-cycle under 50 cc
<Train and Bus>	71	46	Bus, streetcar, and subway
<Route Name>	4	13	Yamanote line, Chuou line, and Nanbu line
<Sky Vehicle>	735	108	UFO and helicopter
<Ship>	988	226	Aegis destroyer and Yacht
(b) Co-occurrence information related to "Facilities"			
<Facilities>	1139	615	Shrine, home, and studio
<School>	500	203	Seed School and School
<Subject>	27	70	Departments of Medicine, Engineering and English
<Faculty>	34	12	Faculties of Medicine, Engineering and Pharmacy
(c) Co-occurrence information related to "Musical instruments"			
<Musical Instruments>	86	1	Musical Instruments
<Wind Instrument>	123	37	Saxophone and flute
<Stringed Instrument>	124	37	Harp, organ, and keyboard
<Percussion Instrument>	117	35	Drum, tabor, and xylophone

Co-occurrence information prepared by human begin. The co-occurrence frequency is given at random by the range from 0 to 9.

4.2.2. Judgment method of automatically system terminating

Fig. 2 shows the transition of the stagnation level with every time executing of the system. The stagnation level increase to some degree with all co-occurrence information of the concepts and then decrease except for "Musical Instruments" because new nodes appear and the movement of the word to various nodes of sub-concepts become large, so their classifications become not clear to the tree.

From the transition in Fig. 2 the judgment method of automatically ending evaluation using the stagnation level is applied. The threshold of the stagnation level cannot uniquely decide because the stagnation level based on *co-occurrence information*. Therefore, the definition of an automatic end of evaluation is

Definition 4.2. Evaluation terminated when the stagnation level decreases than 5% or reaches 100%.

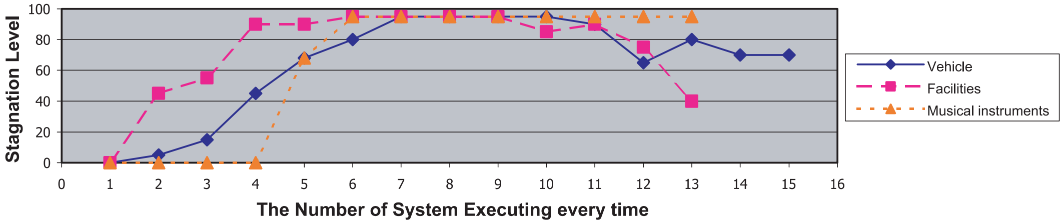

Fig. 2. Transition of stagnation level.

According to Fig. 2 and Definition 4.2 the execution is terminated after 11th times for the concept "Vehicle" and the stagnation level becomes 94.4%. However, the stagnation level is high this not mean that the classification tree is classified correctly. Therefore, it is necessary to examine the constructed system closely to confirm the effectiveness of the stagnation level.

4.3. Experiment method

4.3.1. Reserve knowledge

In the actual experiment, it used by the preliminary experiment in the foregoing paragraph. *Co-occurrence information* related to "Facilities" and "Vehicle" is given roughly the rudders as an advanced knowledge as shown in Fig. 3.

In Fig. 3 the "Vehicle" is summarized all concepts in Table 1(a) and "Facilities" is summarized all concepts in Table 1(b). Moreover, the *receiving words* "gets on", "possible to get on", and "can get on" for "Vehicle" appear 10 times, Also, *receiving words*: "build", "possible to build" and "can build" for word "Facilities" appear 10 times.

The meaning to give the advance knowledge is in the confirmation whether or not is possible to classify its accurately. In the actual experiment, when adding *co-occurrence information* the hierarchy of the sub-concept to the super-concept "Vehicle" and "Facilities" is classified with the common concepts. Moreover, the word with a different meaning is not a sub-concept of these super-concepts, and the tree is constructed. After this advance knowledge is given, the thesaurus tree is classified in the actual experiment.

4.3.2. Content of experiments

In the actual experiment, to verify the efficiency of the new method, about (one million pairs) of co-occurrence information from a data set of 25 Newsgroups from *CNN Web Site* (1995–2002)

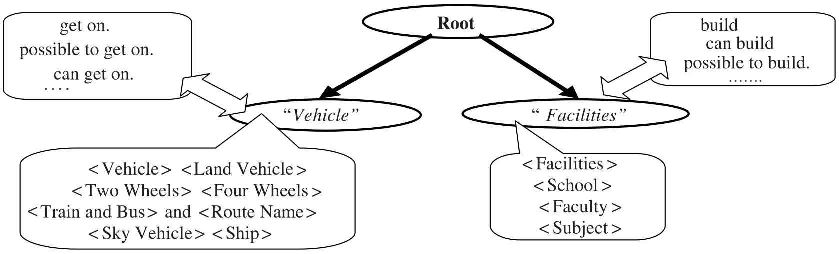

Fig. 3. Preliminary knowledge.

are used. There were various topics related to *sports*, *computers*, *politics*, *economics*, etc. As an experiment method, the above-mentioned *co-occurrence information* is added from the preliminary knowledge system of Fig. 3.

4.4. Experiment result and consideration

The changing in the system by "Vehicle" relation is shown in Figs. 4–6 as a preliminary experiment results. Fig. 4 shows the construction systems in the first time evaluation, Fig. 5 shows the second evaluation and Fig. 6 shows the construction systems after 11th times evaluation. The table that accompanies each node of the concept of the word and the frequency were shown from Figs. 4–6. Sign A to M are used as identification of nodes and assumed to be common. Moreover, in Fig. 6 each node of character string written inside is the one that classified word groups.

From Table 1(a) the relation between the concept of the word in the high rank (super-concept) and the subordinate position (sub-concept) is clear. Therefore, Fig. 6 shows the system made by

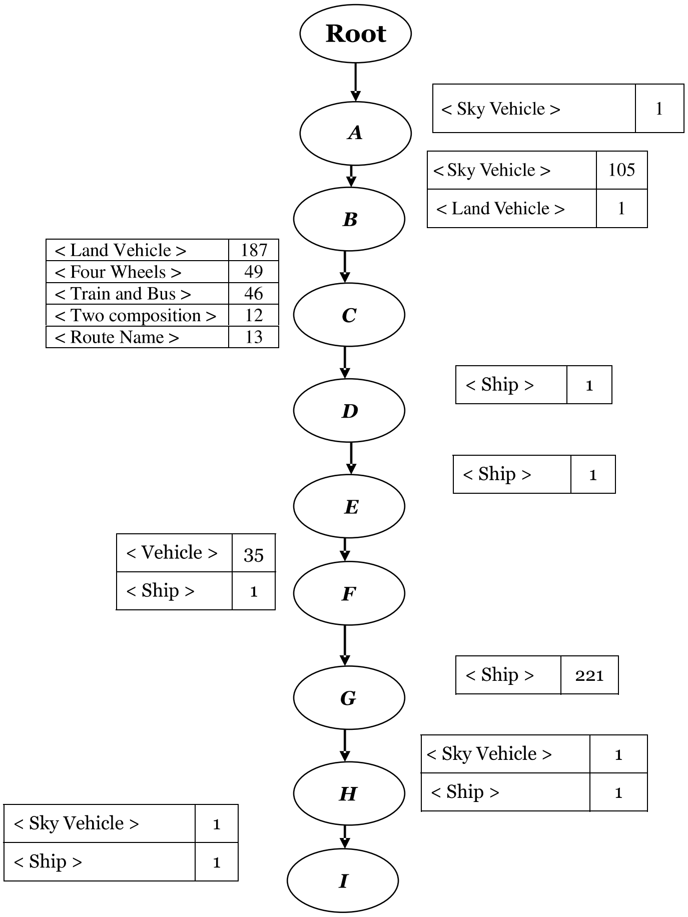

Fig. 4. Construction tree after the first time of executing.

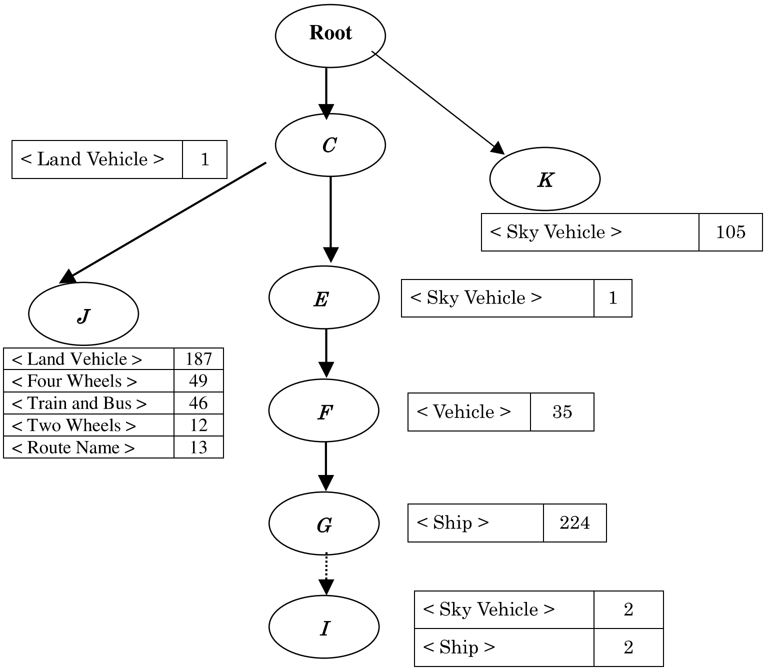

Fig. 5. Construction tree after the second times.

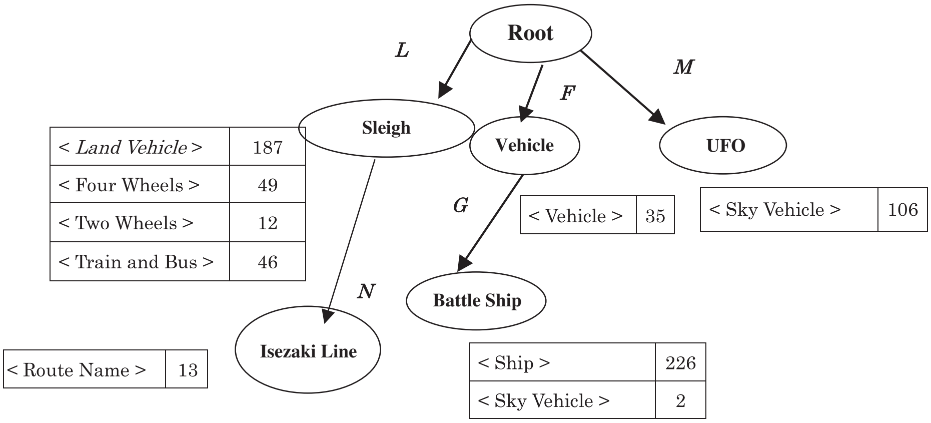

Fig. 6. Construction tree after 11th times.

hierarchically considering eight concepts related to “Vehicle”. Thus, Fig. 7 is the advanced one which having the perfect classification tree. The evaluation is repeatedly by individual nodes that has distributed and then these nodes are settled as shown from Figs. 4–6. Fig. 6 can confirm the change in the system by 11 times because study is ended using Definition 4.2.

In each node of Fig. 6, node *M* represents the concept “Sky Vehicle” and node *G* represents the concept “Ship” etc. The word group with the meaning concentrates respectively is settled. It can

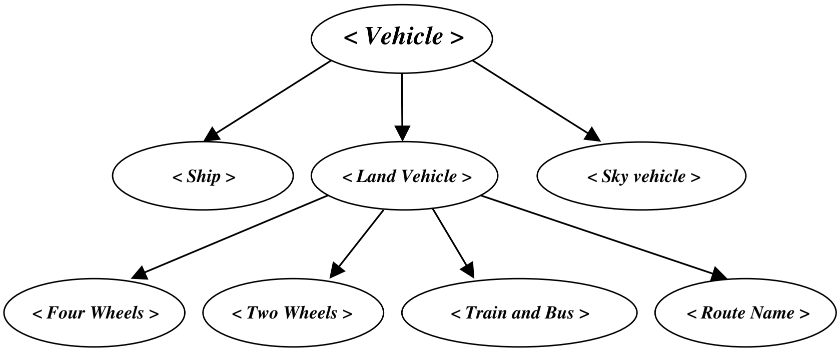

Fig. 7. Perfect classification tree.

be said that it will be classified accuracy. Two words that belong to the concept "Sky Vehicle" have been absorbed to node *G*. These words are "Hovercraft", "Spaceship" which also belongs to both concepts "Sky Vehicle" and "Ship". Therefore, it is also possible to be absorbed them to node *M* and finally they absorbed to node *G* at the 11th times of the evaluation. Moreover, node *L* is settled to one big group not being subdivided because the common characteristics of these three concepts "Four Wheels", "Two Wheels" and "Train and Bus" are settled to the concept "Land Vehicle" even the features numbers of these three concepts are less than the number of features of other concepts as shown in Table 1(a).

However, the word group that belongs to the concept "Route Name" is independently forming that group because the amount of its feature is small. This cause the features in the concept related to other "Vehicle" are gotten without the concept of "Route Name". Moreover, a top to bottom relation of concept "Land Vehicle" and "Route Name" is correct as compared with the correct answer system of Fig. 7.

Overall, the words are classified accurate and a perfect tree is constructed but with some differences in the super-concept and sub-concept relation. Moreover, the system execution is steady from the 7th times but the maintenance of that system is confirmed at the 11th times as in Fig. 6. Therefore, the stagnation level for judging an automatic end of study using Definition 4.2 is effective.

4.5. Evaluations

A part of the system constructed with the actual experiment is separately shown in Fig. 8 after nine times of evaluation. The stagnation level is 54.2%. Fig. 8 gets questions classified as accurately as sub-concept nodes as understood. For example, words related to "Baseball" have gathered in node *A* and words related to "Man" have gathered in node *D*. Moreover, the relation between node *E* and node *F* is understood easily. In a top and bottom relation of the whole, there are a few senses of compatibility group from similarity like the relation between node *B* and node *C* seems to be high and is located in a short distance.

The stagnation level studied in the actual experiment ends at 54.2%. Therefore, it is possible to understand for human even if study is automatically terminated by Definition 4.2

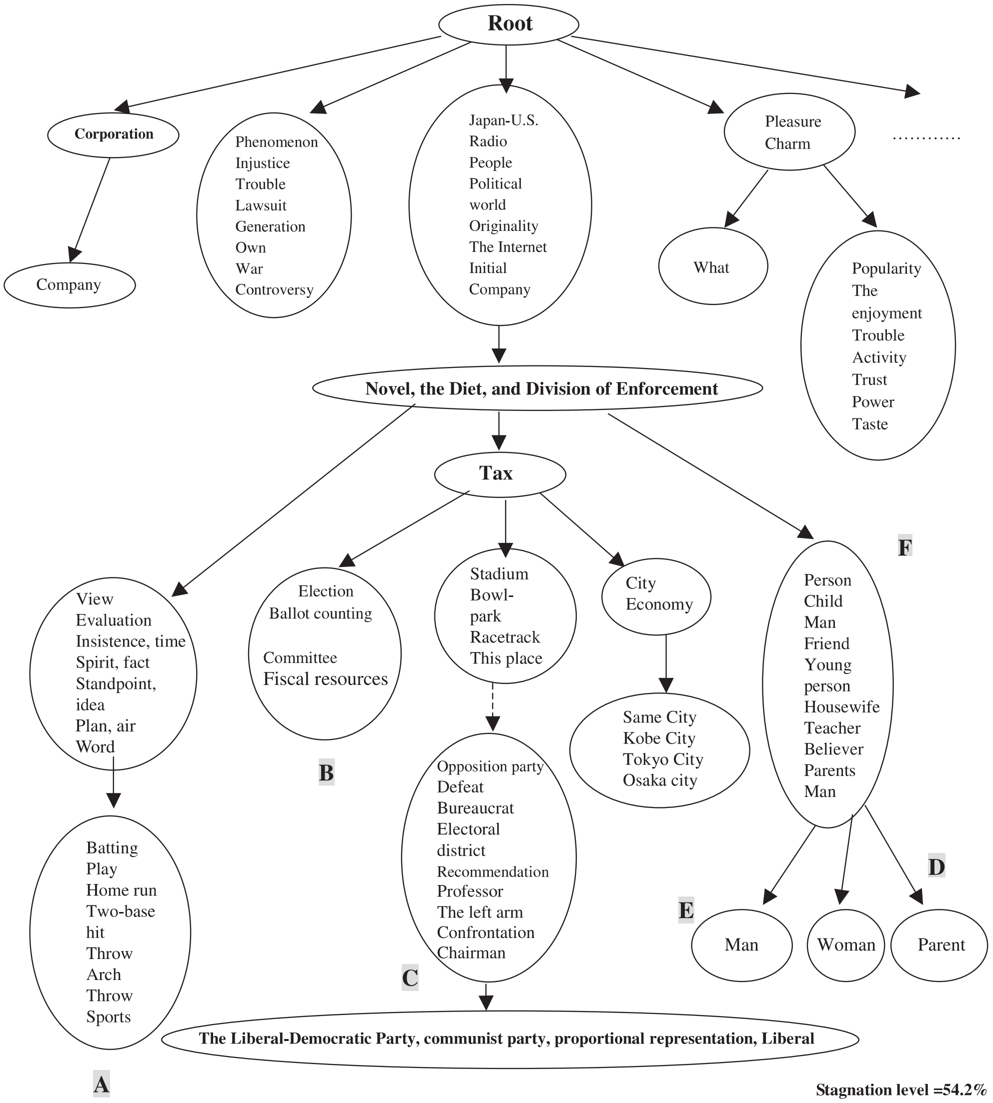
Root
Corporation
Company
Phenomenon Injustice Trouble Lawsuit Generation Own War Controversy
Japan-U.S. Radio People Political world Originality The Internet Initial Company
Pleasure Charm
....
What
Popularity The enjoyment Trouble Activity Trust Power Taste
Novel, the Diet, and Division of Enforcement
Tax
F
View Evaluation Insistence, time Spirit, fact Standpoint, idea Plan, air Word
Election Ballot counting Committee Fiscal resources
Stadium Bowl-park Racetrack This place
City Economy
Person Child Man Friend Young person Housewife Teacher Believer Parents Man
Same City Kobe City Tokyo City Osaka city
B
Opposition party Defeat Bureaucrat Electoral district Recommendation Professor The left arm Confrontation Chairman
Batting Play Home run Two-base hit Throw Arch Throw Sports
D
E
Man
Woman
Parent
C
The Liberal-Democratic Party, communist party, proportional representation, Liberal
A
Stagnation level =54.2%

Fig. 8. Tree constructed by new method.

that the system with knowledge cannot be necessarily constructed. The factor that a system near to the correct answer system could be constructed with the preliminary experiment because the stagnation level is high. However, the result of the actual experiment described in this paper not bad because the limit of the classification according to *co-occurrence information* is used. Finally, in the actual experiment a rough limited classification can been done because the number of features of each word is limited. Therefore, it will be necessary to aim a clustering that uses larger-scale *co-occurrence information* in the future. Moreover, *co-occurrence information* is originally added by a couple, so ending the system construction automatically based on the condition of Definition 4.2 and the complete thesaurus have been constructed.

5. Conclusion

In this paper, the importance of *co-occurrence information* in the NLP system, the classification technique of the co-occurrence word and *co-occurrence frequency* are described. The classified group expressed hierarchically and proposed the technique for constructing the system. Experimental test operation of this system and effectiveness of the proposal technique are examined.

In the future work, *co-occurrence information* should improve the system aiming by adding more each couple and use *co-occurrence information* on a large scale. Moreover, it is necessary to make automatic acquisition of the corpus by using the search engine on *Web* to prepare *co-occurrence information* on a large scale. Also, it is necessary to design the speed-up of the system construction by increasing the number of nodes.

References

Atlam, E.-S., Fuketa, M., Morita, K., & Aoe, J. (2000). Similarity measurement using term negative weight to word similarity. *Information Processing & Management Journal, 36*, 717–736.

Atlam, E.-S., Fuketa, M., Morita, K., & Aoe, J. (in press). Document similarity measurement using field association term. *Information Processing* & *Management Journal.*

Fuketa, M., Lee, S., Tsuji, T., Okada, M., & Aoe, J. (2000). A document classification method by using field association words. *Information Science Journal, 126*, 57–70.

Fukumoto, F., & Tsuji, J. (1994). *Cancellation of polysemy of verb based on corpus*, NLC94-24. Electronic Information Communication Society technology research report (pp. 15–22).

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 29th annual meeting of the Association for Computational Linguistics* (pp. 229–236).

Hirao, K., & Matsumoto, Y. (1994). *Case frame acquisition of verb from corpus and clustering of noun*, NL104-11. Information Processing Society of Japan research report (pp. 79–86).

Kobayashi, Y., Tokunaga, K., & Tanaka, H. (1996). Analysis of compound noun that uses meaning co-occurrence information between nouns. *Natural Language Processing, 3*(1), 29–43.

Koyama, M., & Aoe, J. (1995). High speed searching algorithm of hierarchy concept dictionary. In *The 51st national athletic meeting, Vol. 7 E-2* (pp. 4-235–4-236).

Li, H., & Abe, N. (1997). Clustering words with the MDL principle. *Journal of Natural Language Processing, 4*(2), 71–87.

Matsumoto, L. (1992). Rule related to the co-occurrence between words that use analysis tree data base. *Electronic Information Communication Society Thesis Magazine, J75-D-5*(3), 589–600.

Morimoto, K., Iriguchi, H., & Aoe, J. (1993). A retrieval algorithm of dictionaries by using two trie structures. *Transaaction of the IEICE, J76-D-II*(11), 2374–2383, in Japanese.

Morita, K., Mochizuki, H., Yoshihiro, Y., & Aoe, J. (1998). Efficient retrieval algorithm of co-occurrence information that uses trie structure. *Information Processing Society of Japan Thesis Magazine, 39*(9), 2563–2571.

Petira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In *Proceedings of the 31st annual meeting of the Association for Computational Linguistics* (pp. 183–190).

Takahashi, N., & Itabashi, S. (1998). *Morphological analysis using word co-occurrence frequency*, 88-NL-69. Information Processing Society of Japan Natural Language Processing Society (pp. 1–8).

Tokunaga, T., Iwayama, M., & Tanaka, H. (1995). Automatic thesaurus construction based-on grammatical relations. In *Proceedings of IJCAI'95*.

Yokoyama, H., & Shinichiro, O. (1998). *Classification of meaning of Japanese verb that uses co-occurrence information from corpus*, NLC97-55. Electronic Information Communication Society (pp. 1–8).

Zhang, J., & Rasmussen, E. (2001). Developing a new similarity measure from two different perspectives. *Information Processing & Management Journal, 37*(2), 279–294.