Extraction of Associative Attributes from Nouns and Quantitative Expression of Prototype Concept

Maya Ando, Jun Okamoto and Shun Ishizaki

Graduate School of Media and Governance, Keio University
5322 Endo, Fujisawa-shi, Kanagawa, 252-8520, Japan
{maya, juno, ishizaki}@sfc.keio.ac.jp

Abstract

One of the purposes of this research is to formalize similarity among nouns by using attributes associated from the nouns, and then using the similarity, to formalize prototypes of categories. The other purpose is to extract features of nouns by using adjectives or adjective-like words obtained by the association experiments and to formalize importance of the nouns with the words.

We constructed an associative concept dictionary using many kinds of attributes associated from nouns. Similarity among nouns was calculated by using their associated attributes with inner product methods, where the nouns were organized in a hierarchical structure using generalized or specific relations.

This paper discusses similarity between nouns using their attributes. We found that the similarity of nouns located at lower levels has a high score in many cases. Then prototypes are quantitatively formalized among Japanese noun concepts. It uses similarities of part/material concepts, features, and action concepts, and distance values between the noun and its lower-level concepts. Such formalized prototypes are compared with a result of human questionnaire experiments to obtain a good correspondence among them.

1. Introduction

One of the purposes of this research is to formalize prototype concepts of a noun by using attributes associated from the noun. We extracted various attributes from nouns by associative experiments and calculated similarity between the nouns and its lower-level concepts. In this paper we formalized an expression of prototype concepts (Rosch,1975) using the similarities. The prototype concepts were verified quantitatively by comparing them with results of questionnaires asking human subjects about prototype concepts from the nouns. The other purpose is to extract concepts representing features of nouns by adjectives or adjective-like words by the associative experiments and to formalize importance values with them(Ando,1999). Results obtained from the research will be applied to information retrieval systems, flexible question-answering systems, or automatic construction systems for concept dictionaries.

Nouns have various attributes as well as their word senses written in linguistic dictionaries. For example, a concept of fruit has attributes such as “food”, “has-seed”, “has-vitamin c” or “delicious”. One of the ways to obtain such attributes is an associative experiment with humans(Okamoto,2001a).

Large-scale association experiments were done to extract attributes from nouns. Stimulus words were consisting of basic nouns that were selected from textbooks used at Japanese elementary schools. Association was done from stimulus words with 7 tasks, generalized concepts, lower-level concepts, part/material concepts, features, synonyms, action concepts and situation concepts.

Similarity was calculated between nouns by using their associated attributes with inner product methods. The nouns were connected in a hierarchical structure using generalized or specific relations obtained by using results of the association experiments. This paper discusses similarity between nouns using their attributes along the hierarchical structure as well as the prototype of Japanese noun concepts. We found that the similarity of attributes of nouns located at lower levels has high score in many cases. Using the similarities, prototype of a concept can be formalized in order to find typical concepts among its lower-level concepts. The results have good correspondence with questionnaires with human subjects.

2. Associative Concept Dictionary

G. H. Kent and A. J. Rosanoff reported the large-scale study of associative experiments in 1910(Miller,1991). They read aloud a list of 100 stimulus words to a subject who was instructed to give “the first word that occurs to you other than the stimulus word.” The number of the subjects is 1000 men and women and they have different occupations and level of education. As a result most of the associated words can be classified as instances of just four kinds of connotative semantic relations:

1. Super-ordinate, coordinate, and subordinate terms: terms that arrange things in a taxonomic tree
2. Attributive terms: modifying terms that state the values of attributes of things
3. Part –whole relation: terms that name a part of something, or that name the whole of which something is a part
4. Functional terms: terms that designate the ends that things serve – what things normally do or what is normally done with them

We carried out associative experiments for the purpose of constructing an associative concept dictionary. The association experiments consist of a set of a stimulus noun and seven tasks. In the total of 1,000 stimulus nouns, about 800 are chosen from "basic nouns" in Japanese elementary school textbooks. 200 of the nouns were chosen from basic nouns obtained through the association experiments.

The 7 tasks are generalized concepts, lower-level concepts, part/material concepts, features, synonyms, action-concepts and situation concepts in each stimulus noun.

Current total number of associated words is more than one hundred and twenty three thousand.

Subjects include undergraduate students and master course students at Keio University. Number of subjects for a stimulus word is fifty for this research. Some of the stimulus words have ten subjects but the subject number is now increasing to fifty.

The association experiment is carried out on the computer network in Keio University, Shonan Fujisawa Campus. The procedure of the experiment is as follows. A subject associates to a stimulus word and inputs associated words into the system using a word processor. Then the system calculates an association time, an order of an associated words, and frequency of associated words.

Distance between a stimulus word and its associated word is calculated quantitatively using a linear programming method.

Let the distance, D, between a stimulus word and an associated word be shown in the following linear function:

$$D=\alpha F+\beta S+\gamma T ,$$

where 3 parameters T, S and F are defined by

$$F=\frac{N}{n+\delta},\ \delta=\frac{N}{10}-1 ,$$

$$S=\frac{1}{n}\sum si\ ,\ T=\frac{1}{n}\sum\frac{t_i}{60}\ .$$

In these parameters, ti is a time taken for the association by a subject, si is an order of the association by a subject, N is a number of the subjects, n is a number of subjects who input a same associated word from a stimulus word at each semantic relation, and δ is a value of modification. We need δ because the number of the subjects is not fixed and ranged from 10 to 50.

The boundary conditions in the linear programming method are defined in the following two cases.

1."all of the subjects associate the word", "the word is associated at first" and "response time is short"

2."the number of subject that associate the word is only one", "an association order is low" and "response time is considerably long"

By using the Simplex method, α , β and γ were calculated. We found the third parameter to be 0.

The following result is obtained by using the linear programming method:

$$D=0.81F+0.27S.$$

The associative concept dictionary includes stimulus noun words, tasks, frequency of association, order of association, duration of association and distance of the stimulus word and associated words. In this paper, we use the attributes associated by 50 subjects and examine the case that frequency of association is more than 0.4. It means that the attribute is associated by more than 2 subjects.

```
(vegetable
  (generalized concept
    (plant          0.21   1.346  0.52   1.713)
    (food           0.236  1.28   0.5    1.742) )
  (specific concept
    (carrot         0.321  2.5    0.56   1.941)
    (tomato         0.411  3.13   0.46   2.345) )
  (part/material
    (root           0.354  2.5    0.36   2.516)
    (leaf           0.349  2.467  0.3    2.798)
    (seed           0.229  2.125  0.16   3.949) )
  (feature
    (delicious      0.292  1.286  0.28   2.597)
    (green          0.432  1.182  0.22   3.019)
    (cheap          0.383  2      0.02   8.64) )
  (synonym
    (greens         0.55   1      0.02   8.37) )
  (action-related
    (eat            0.174  1.457  0.7    1.432)
    (cut            0.192  2.385  0.26   3.026)
    (sell           0.393  3.8    0.1    5.526) )
(situation-included
    (field          0.191  1.632  0.76   1.405)
    (fruit and vegetable shop
                    0.29   2.269  0.52   1.963)
    (kitchen        0.432  3.077  0.26   3.213) ) )
```

Figure 1. An Example of The Concept Description for Stimulus Word "vegetable"

3. Formalization of Similarity between Nouns

3.1. Formalization of Similarity

The similarity is formalized using associated words with their distance values from the stimulus word. We define similarity, S between a stimulus word A and a stimulus word B, as follows.

Table 1: Example Matrix for Similarity Formalization

	Stimulus Word A	**Stimulus Word B**
Associated Word 1	1.35	2.18
Associated Word 2	3.73	0
Associated Word 3	4.07	4.31
:	:	:
Associated Word *i*	*aj*	*bj*
:	:	:

$$X_i=\frac{\frac{1}{a_i}}{\sum(\frac{1}{a_j})}\qquad Y_i=\frac{\frac{1}{b_i}}{\sum(\frac{1}{b_j})}$$

$$S = \frac{X \cdot Y}{|X| \cdot |Y|}$$

where *aj* is a distance value between a stimulus word A and an associated word and *bj* is a distance value between stimulus word B and an associated word. The reciprocal of the distance value is used for normalization, because the distance value becomes more important when it decreases. Similarity S is shown as an inner products. To normalize the similarity value S, S is divided by norms of vectors X and Y. When two words are same, the inner product is 1, and if they are similar, it is close to 1.

We calculated similarity values of stimulus nouns in a hierarchal structure. Attributes of seven tasks are used for the similarity calculation. The results are shown in Table2. It shows that part/material concepts, features and action-concepts have high similarities. Table 2 is an example of the comparison of similarity among "n*orimono"* and its associated words. "Norimono" is a generalized concept of vehicle, vessel and airplane. *"Norimono*' has specific concepts such as "car", "train" and "airplane." In the table, we consider high similarity with values of more than 0.5.

In this paper, we will use the similarities of these three tasks, lower-level concepts, features and action concepts.

Table 2: Similarity about 7 tasks

Generalized concept	0.224
Lower-level Concept	0.044
Part/Material	0.583
Feature	0.584
Synonym	0
Action-related	0.629
Situation-included	0.382

4. Analysis of Similarity Tendency by Using Inheritance of Attributes

As for the inheritance of attributes in a hierarchical structure, it is generally said that attributes in higher level concepts are inherited to those of lower level concept. In this paper, using nouns included in three hierarchical levels, which means each nouns included in three successive levels, similarity among the nouns is calculated.

We examined four big categories *"kudamono*[fruit]*"*, *"norimono*[the generalized concept including vehicle, vessel and airplane]*"* , *"gakki*[musical instrument]*"* and *"kagu*[furniture]*"*. The number of associated words depends on tasks or the stimulus nouns themselves. We selected the four different types of categories for this research based on average numbers of associated words with a task of lower level concepts.

"N*orimono*" has many associated words in a task of part/material concepts. "Fruit" has many associated words in action concepts. "Furniture" and "musical instrument" have little difference on the number of the associated words.

For the sake of convenience, we called the similarity between the highest level concept and the middle level concept as Similarity A. The similarity between the middle level concept and the lowest level concept is Similarity B. Finally, the similarity between the highest level concept and the lowest level concept is Similarity C.

Table 3: Similarity among "fruit-grape-muscat"

	fruit/grape	grape/muscat	fruit/muscat
Part/Material	0.74	0.88	0.76
Feature	0.67	0.77	0.77
Action-related	0.68	0.71	0.75
average	0.7	0.79	0.76

Table 4: Similarity among "norimono-train-subway"

	norimono/train	train/subway	norimono /subway
Part/Material	0.72	0.77	0.56
Feature	0.75	0.68	0.53
Action-related	0.79	0.87	0.78
average	0.75	0.77	0.62

Table 5: Similarity among "musical instrument-piano-grand piano"

	instrument/ piano	piano/ grand piano	instrument /grand piano
Part/Material	0.59	0.96	0.66
Feature	0.53	0.82	0.46
Action-related	0.73	0.89	0.66
average	0.62	0.89	0.59

Table 6: Similarity among "furniture-chair-rocking chair"

	furniture/ chair	chair/ rocking chair	furniture/ rocking chair
Part/Material	0.81	0.86	0.80
Feature	0.48	0.57	0.33
Action-related	0.49	0.58	0.23
average	0.59	0.67	0.45

We found that stronger inheritance of attributes appeared in lower level concepts because Similarity B has larger values than those of Similarity A. We also found Similarity A is larger than Similarity C.

We found two characteristics of this similarity inheritance. One is that both the averages of similarity A and B have high similarity values such as "norimono" and "fruit". Similarity A of "norimono" and that of "fruit" are more than 0.7 and Similarity B of these categories has also a high value. On the other hand, the similarity A of categories whose lower level concepts have a variety of features, such as shapes or usages of "furniture" and "musical instrument," is less than about 0.6 and Similarity B has higher values than those of Similarity A. As a result we found that the inheritances of attributes has a few types depending on categories.

5. Prototype of Category

One of words in a category can be a prototype and others have less prototypeness.. Eleanor Rosch discussed in proposing her prototype theory by psychological experiments(Rosch, 1975). For example, a bird has lower level concepts such as pigeons or penguins. Pigeons are more typical than penguins because pigeons can fly but penguins cannot.

5.1. Questionnaire Experiment of Prototype

At first, we thought that the words with low distance values in the associative concept dictionary would be prototypes. After watching the dictionary, we found that this is not the case. To get a perspective of prototypes, we made questionnaire experiments. Requests to the subjects to get prototypes for a concept were "please list up three words just like the given concepts."

Prototypes of categories depend on subject's situation. For example, robins are familiar in the US and can be typical birds. On the other hand, in Japan robins are not familiar and cannot be a prototype. In the questionnaire experiments, subjects are students at Keio University. The result of our experiments will show prototypes of young Japanese people.

We gave points to the words listed up by the subjects. The point to a word listed at the first position is 3, the second word 2, and the third 1.We arranged the words according to the points by summing up the point with each subject. As a result we find that the questionnaire results differ from that depending on the low distance values of lower-level concepts. In the case of "furniture," Table 7 shows that the order of lower-level concept differs from the order of questionnaire. In the case of "fruit," Table 8 shows that the order of lower-level concept and the order of questionnaire are almost same.

We calculated the similarity about part/material concepts, features, and action concepts in associative concept dictionary to extract crucial information to define the prototypes. We found that when a word is typical one, its score of similarity is high. But these kinds of information was not enough to define prototype of a category.

Table 7:Comparison between Lower-level Concept and Questionnaire "Furniture"

	Distance of lower-level concept(order)	score of questionnaire(order)
Chair	1.47(1)	26(2)
chest of drawers	1.62(2)	39(1)
Desk	1.99(3)	17(3)
Table	2.95(4)	7(5)
Bed	3.41(5)	9(4)

Table 8:Comparison between Lower-level Concept and Questionnaire "Fruit"

	distance of lower-level concept (order)	score of questionnaire (order)
Apple	1.59(1)	41(1)
Orange	1.99(2)	29(2)
strawberry	3.09(3)	15(3)
Grape	3.12(4)	2(5)
banana	3.21(5)	12(4)

5.2. Formalization of Prototype of Category

In the last chapter, we discussed inheritance between concepts included in three hierarchical levels. We showed that similarities between highest level concepts and lowest level ones have larger values. In the case of "furniture", the questionnaire result shows that "bookshelf" is not prototype of "furniture", but similarity between "furniture" and "bookshelf" has large values.

Next, we decided to use distance values of lower-level concepts as well as similarities of part/material concepts, features, and action concepts. We defined prototype P as a linear combination of three similarity values divided by the distance value of the lower level concept. The formalization means that the lower distance values are more important than that of larger values.

$$P=(\alpha M+\beta F+\gamma V)/D ,$$

M: similarity of part/material concept
F: similarity of feature
V: similarity of action concept
D: distance of lower level concept
α , β , γ are linear coefficients

As the first approximation, we simply set the coefficients as 1.0. Then, we arranged words according to the P values.

By using spearman's rank correlation coefficient(Ikeda,1976), it is shown that the order of P and the order in the questionnaire correspond each other. The order of questionnaire and the sum of three similarity are also corresponding.

We found that the correlation between the questionnaire results and P has larger values for any categories. It can be said that prototype of the category is represented by using distances of lower level concepts and similarities of the three tasks.

Table 9:Data of "furniture"

	order of lower-level concept	order of questionnaire	similarity of part/ material	Similarity of feature	similarity of action concept	P (order)
chair	1	2	0.81	0.48	0.49	1.21(2)
chest of drawers	2	1	0.78	0.73	0.64	1.32(1)
desk	3	3	0.87	0.64	0.52	1.02(3)
table	4	5	0.71	0.63	0.54	0.64(4)
bed	5	4	0.53	0.55	0.31	0.41(5)
sofa	6	6	0.37	0.27	0.44	0.25(9)
bookshelf	7	7	0.76	0.85	0.55	0.45(7)
shelf	8	8	0.82	0.65	0.57	0.40(6)
cupboard	9	9	0.75	0.61	0.58	0.32(9)

Table 10:Rank Correlation Coefficient with questionnaire result

	P	Total of three similarities
furniture	1	0.7
fruit	0.8	0.31
norimono	0.98	0.92
musical instrument	0.90	0.6

6. Conclusion

We showed various attributes of nouns can be used to construct an inheritance structure obtained from an associative concept dictionary. Using lower level concepts, part/material concepts, features and action concepts, we can define prototype P. We showed the higher value of the correlation between questionnaire and the P of prototype to verify the effectiveness of the formalization.

As a future task we need to improve P formalization because all the coefficients are set to 1..0 as the first approximation in this paper. The coefficients are to be calculated aiming at more precise results.

The association experiments will continue because we need to increase subjects and stimulus words to apply the concept dictionary .to QA systems, IR systems or automatic ML systems.

7. References

George A. Miller, 1991. *The science of words*, chapter 8. Scientific American Library.

E. Rosch, 1975. *Cognitive representations of semantic categories*, Journal of Experimental Psychology General 104:192-233.

Kai, M. and Matsukawa, T, eds, 1996. *Goi shidou no houhou [Method for Guidance of Vocabulary]*. Mitsumura Tosho.

Okamoto, Jun. and Ishizaki, Shun, 2001. *Construction of associative concept dictionary with distance information, and comparison with electronic concept dictionary*. Journal of Natural Language Processing (in Japanese), vol.8 No.4:37-54.

Okamoto, Jun. and Ishizaki, Shun, 2001. *Associative concept dictionary construction and its comparison with electronic concept dictionary*. Pacific association for computational linguistics 2001.

Ando, Maya. Okamoto, Jun. and Ishizaki, Shun., 2000. *Feature of associated attributes and quantitative research.* Special inter group on language sense processing engineering.

Ikeda, T, 1976. *Toukeiteki houhou 1[Statistical Method 1]*, Shinyousha.