# Associative Concept Dictionary Construction and its Comparison with Electronic Concept Dictionaries

Jun Okamoto and Shun Ishizaki
Keio University
Graduate School of Media and Governance
5332 Endo, Fujisawa-shi, Kanagawa, 252-8520, Japan
TEL: +81-466-48-6101
E-mail: juno@sfc.keio.ac.jp, ishizaki@sfc.keio.ac.jp

**Summary**

An associative concept dictionary is constructed using large-scale online association experiments. In the experiments, subjects are presented stimulus words chosen from elementary school textbook of Japanese. They are requested to make association from the stimulus words about 7 semantic relations, such as, higher-level concepts, lower-level concepts, actions, situations. The dictionary is constructed by using all of the associated concepts, which are connected to the stimulus words with calculated distances. The distances are obtained using a linear programming method, which combines two parameters linearly, the frequency of the associated word and the associated order of the word. The associative concept dictionary is compared with the conventional concept dictionaries such as EDR concept dictionary, Nihongo Goi-Taikei and WordNet. The comparison shows that the associative concept dictionary is more similar to WordNet than EDR dictionary or Goi-Taikei.

**Key Words:** Association Experiment, Linear Programming, Distance between Concepts, Electronic Concept Dictionary, Quantification of Concept Space

## 1. Introduction

Background knowledge concerning input texts is necessary when a computer tries to understand the contents of the text as well as the syntactic and semantic information of it. In this research, the associative concept dictionary is built based on the concepts obtained by using large-scale online association experiments. The dictionary does not only include semantic and contextual information about the stimulus words but also conceptual hierarchy information. The conventional concept dictionaries have tree structures for the hierarchy. Their distances between concepts in the dictionary are calculated using the number of links between them, whereas the associative concept dictionary has quantitative distance information between two concepts, which is calculated by using a linear programming method. The method combines two parameters linearly, the frequency of the associated word and the associated order of the word in the association experiment. This paper shows comparison among the associative concept dictionary, EDR concept dictionary (EDR 1990) (hereafter EDR), Nihongo Goi-Taikei (Ikehara et al. 1997) (hereafter Goi-Taikei) and WordNet (Miller et al. 1993) with the distance information using Principal Component Analysis. Familiar words are chosen as stimulus words in order to compare the four dictionaries.

- The case that the distance, D2, is supposed to satisfy the following conditions, “the number of subject who associate the word is only one”, “association order is lower” and “response time is considerably long”

By using the simplex method, $\alpha$, $\beta$ and $\gamma$ are calculated. We found the first two parameters significant for the distance calculation and the third parameter to be zero, which mean unnecessary of the parameter.

The following result are obtained in the above formula (1):

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

### 2.3. Construction of Associative Concept Dictionary

Using the quantified distances, the associative concept dictionary is built by organizing the stimulus concepts and their associated concepts. The dictionary is organized in a hierarchical structure with the higher-level concepts and lower-level concepts, as well as attribute information to explain the feature, synonym of the stimulus word. It also includes action concepts and situation concepts related to the stimulus words. A used-in slot is employed to show stimulus words, which give the stimulus word as an associated concept.

```
(chair                  <1>   <2>   <3>   <4>
  (higher-level concept   ↓     ↓     ↓     ↓
    (furniture          0.92  1.02  0.16  1.09)
    (object             0.04  2.50  0.24  7.43))
  (lower-level concept
    (sofa               0.48  1.92  0.42  1.96)
    (rocking-chair      0.28  1.43  0.59  2.64))
  (part/material
    (wood               0.60  1.20  0.14  1.52))
  (attribute
    (hard               0.46  1.17  0.32  1.82))
  (synonym
    (seat               0.02  1.00  0.15  8.37))
  (action
    (sit down           0.70  1.03  0.15  8.37))
  (situation
    (school             0.30  2.40  0.22  2.78))
  (used-in
    (furniture  lower-level concept)
    (school     part/material)))
```

**Fig. 1 An example of the concept description for "chair" (partial presentation)**

In Fig. 1, "chair" is a stimulus word. And "furniture" is a higher-level concept of "chair". The numbers <1>, <2>, <3> and <4> in Fig. 1 express a frequency of subjects who gave the same associated word, an average of order of association, an average of response time and the conceptual distance, respectively. All the numbers are normalized for comparing each other.

### 2.4. Reliability of Association Experiment

We carried out the association experiment using basic nouns for stimulus words for evaluation of the experiment. 100 subjects per a stimulus word were requested to make association. The stimulus words are "grape", "fruit", "vehicle", "cherry tree" and so on (in Japanese). In the present paper, reliability of the experiment is estimated by the split-half method. This method split at random the data into two sets of a same number. The high correlation between the two halves may show reliability. We calculate the frequency of the associated word at each stimulus word about 7 semantic relations, based on each half of the scales (100 subjects are split at random into two group). Then reliability coefficients are also calculated for each stimulus word. Correlations between these two groups at each stimulus word were high. For example, the correlation of "fruit" is .85, "grape" is .91, and "cherry tree" is .87.

## 3. Comparison with the Other Electronic Concept Dictionaries

### 3.1. EDR

EDR is an electronic concept dictionary, which contains a classification of concepts organized with respect to super-sub (is-a) relation. (EDR 1990; Utiyama and Hashida 1997)

Since EDR is a bilingual (Japanese / English) concept dictionary, it is used for matching of the concepts in WordNet and the associative concept dictionary.

### 3.2. Goi-Taikei

Goi-Taikei is a Japanese – English dictionary of machine translation system, which consists of three parts: the semantic attribute system, the word dictionary, and the valency dictionary. In terms of size, the dictionary is very large containing 400,000 words. The concept hierarchy for common nouns of Goi-Taikei comprises 2,710 classes in a 12-level tree structure (Asanoma 2001, Ikehara et al. 1997).

### 3.3. WordNet

WordNet is an English electronic concept dictionary. It organizes English word and phrases into synonym sets ("synsets") representing underlying lexical concepts. Synsets are linked with each other via relationships such as super-sub and antonym. (Miller et al. 1993; Utiyama and Hasida 1997) Concepts and synsets are equally regarded as sets of words (and phrases) in this paper.

### 3.4. Distance Measurement of Four Dictionaries

We compare the associative concept dictionary with EDR, Goi-Taikei and WordNet using the distance information. Conceptual distance between two concepts is defined as simple link-counting of the tree structure, considering link direction, relative depth and density. (Richardson, R. et al. 1996; Budanisky and Hirst 2001) In this paper, the distances between two concepts $w_1$ and $w_2$ of the four dictionaries are defined as follows:

$$D_{acd}(w_1,w_2) = \min_j \sum_{i=0}^{n_j-1} D_{acd}(c_i^j, c_{i+1}^j),$$

$$D_{EDR}(w_1,w_2) = \min_j \sum_{i=0}^{n_j-1} D_{EDR}(c_i^j, c_{i+1}^j),$$

$$D_{Goi}(w_1,w_2) = \min_j \sum_{i=0}^{n_j-1} D_{Goi}(c_i^j, c_{i+1}^j),$$

$$D_{WN}(w_1,w_2) = \min_j \sum_{i=0}^{n_j-1} D_{WN}(c_i^j, c_{i+1}^j),$$

where $w_1 = c_0^j$, $w_2 = c_{n_j}^j$, $c_i^j$ is an i-th concept on the j-th path from $w_1$ to $w_2$, and $n_j$ is a number of concepts on the j-th path. It is supposed that there are $n_j$ paths from $w_1$ to $w_2$.

The distances $D_{acd}(w_1,w_2)$, $D_{EDR}(w_1,w_2)$, $D_{Goi}(w_1,w_2)$ and $D_{WN}(w_1,w_2)$ are the distance of the shortest path between two concepts $w_1$ and $w_2$ of four dictionaries. For example, $D_{acd}$("*jidousya* - automobile","*kikai* - machine") is 3.07. This is the shortest path via "*norimono* - vehicle".

### 3.5. Comparison using Example of "vehicle"

Figs. 2,3,4, and 5 show the distance between two concepts (higher-level concepts and lower-level concepts) of the four dictionaries: the associative concept dictionary, EDR, Goi-Taikei and WordNet. The lower-level concepts are "*jidousya* - automobile", "*supo-tuka-* - sports car", "*densya* - train" and "*chikatetsu* - subway". The higher-level concepts are "*norimono* - vehicle", "*dougu* - tool", "*kikai* - machine", and "*mono* - object".

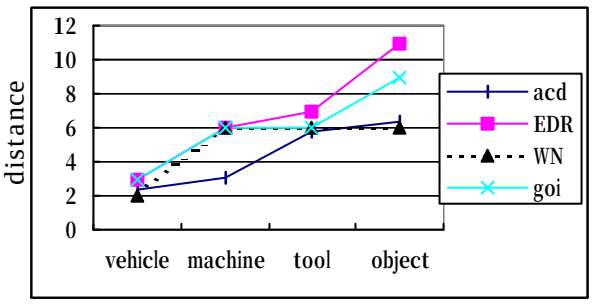


**Fig. 2 Distance between "automobile" and its higher-level concepts**

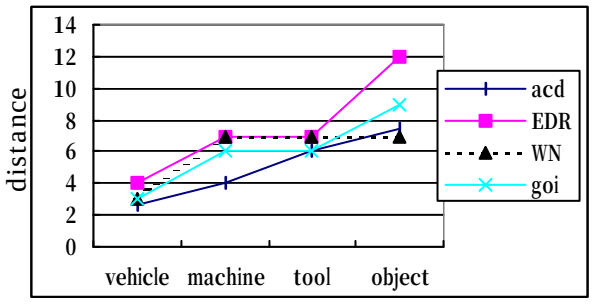


**Fig.3 Distance between "sports car" and its higher-level concepts**

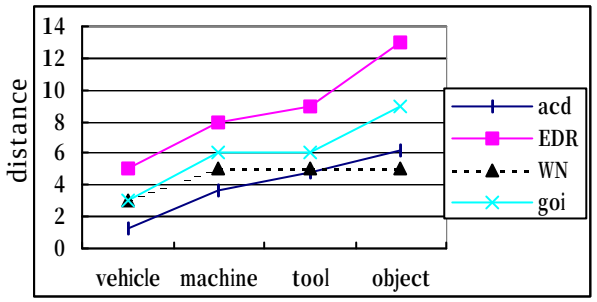


**Fig.4 Distance between "train" and its higher-level concepts**

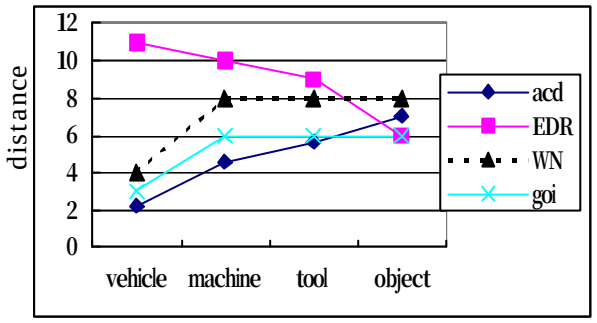


**Fig.5 Distance between "subway" and its higher-level concepts**

In Fig. 2,3 and 4, there is an intermittent increase in the associative concept dictionary, Goi-Taikei and EDR. The distance between two concepts in EDR is longer than the distance between those in the associative concept dictionary and Goi-Taikei. In Fig. 5, an increase is found in associative concept dictionary, Goi-Taikei and WordNet, while the distance between two concepts of EDR gradually decreases. This differs clearly from the other three dictionaries. EDR has "place" and "track" as higher-level concepts of "subway", but does not have "vehicle". Since the shortest path between the two concepts "subway" and "vehicle" passes via "object", the distance becomes much longer.

Fig. 6 shows a principal component analysis of a two-dimensional data cloud about vehicle. Small circles show "automobile", "sports car", "train", and "subway" of each dictionary. The horizontal line shows the first principal component, and the vertical line shows the second principal component. In Table 2, the contribution ratio is shown by the percentage of variance explained by each principal component.

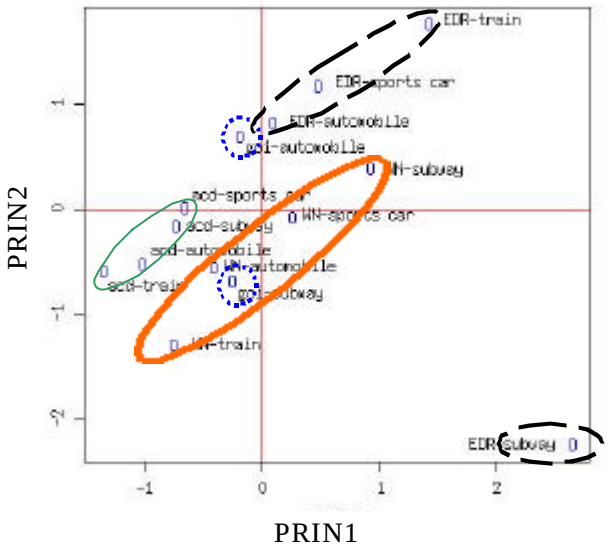


**Fig.6 Principal component analysis of vehicle**

**Table 2: Eigenvalues of the covariance matrix**

| | PRIN1 | PRIN2 | PRIN3 |
|---|---|---|---|
| eigenvalue | 2.63 | 1.12 | 0.14 |
| contribution ratio | 65.83 | 28.06 | 3.50 |

The data cloud of the associative concept dictionary is much closer to WordNet than EDR. The data clouds of EDR and Goi-Taikei are divided into two groups: "subway" and the other concepts. These two evidences clearly show that the conceptual structure of the associative concept dictionary is comparatively similar to WordNet in same parts. On the other hand, the distances between two concepts were comparatively long in EDR, because EDR has middle nodes on the path, which is "function, form, and evaluation" in the concept structure.

### 3.6. Comparison using Example of "plant"

Fig.7,8, and 9 show the distance between two words (lower-level concepts and higher-level concepts) of the four dictionaries: the associative concept dictionary, Goi-Taikei, EDR and WordNet. The lower-level concepts

are "*sakura* - cherry tree", "*budou* - grape", "*kudamono* - fruit". The higher-level concepts are "*syokubutsu* - plant", "*seibutu* - living-thing" and "*mono* - object".

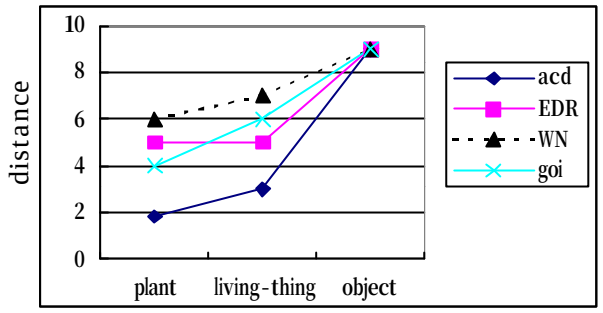


**Fig. 7 Distance between "cherry tree" and its higher-level concepts**

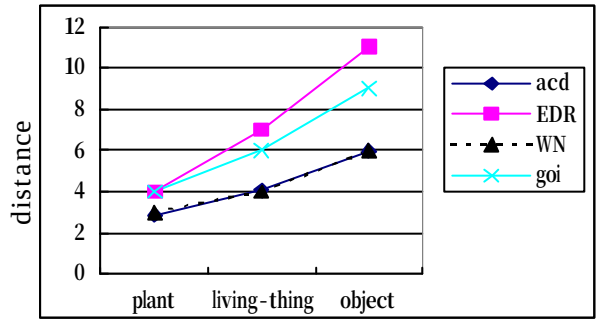


**Fig. 8 Distance between "grape" and its higher-level concepts**

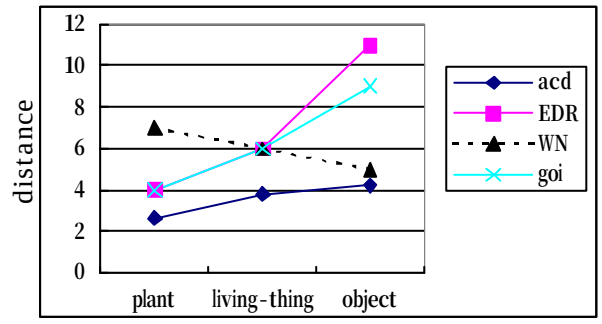


**Fig. 9 Distance between "fruit" and its higher-level concepts**

In Fig.9, there is an increase in the associative concept dictionary, Goi-Taikei and EDR, but the distance between two concepts in WordNet gradually decreases, showing the difference from the other dictionaries. WordNet has "plant part" as the higher-level concept of "fruit". Since the shortest path between two concepts, "fruit" and "plant" passes via "object", the distance is much longer.

Fig. 10 shows a principal component analysis of a two-dimensional data cloud about plant. Small circles show "fruit", "grape", "muscat", "vegetable", "carrot", "spinach", and "cherry tree" of each dictionary. The horizontal line shows the first principal component, and the vertical line shows the second principal component. In Table 3, the contribution ratio is shown by the percentage of variance explained by each principal component.

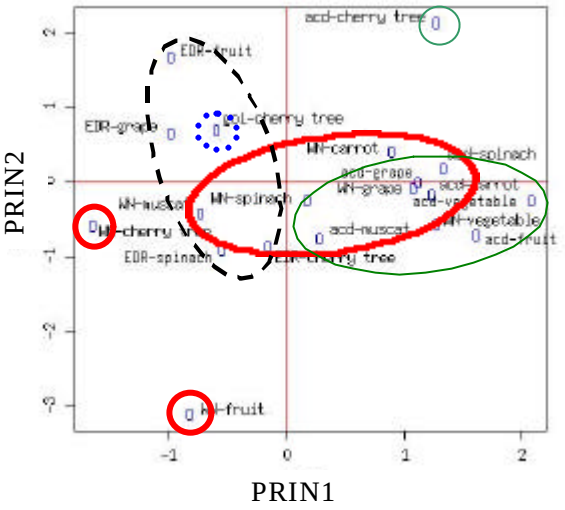


**Fig.10 Principal component analysis of "plant"**

**Table 3: Eigenvalues of the covariance matrix**

| | PRIN1 | PRIN2 | PRIN3 |
|---|---|---|---|
| eigenvalue | 2.11 | 0.78 | 0.10 |
| contribution ratio | 70.47 | 26.08 | 3.45 |

The data cloud of the associative concept dictionary is divided into two groups: "cherry tree" and other concepts. There is a plausible reason why the distances from "cherry tree" to higher-level concepts are rather short compared with the other dictionaries. Because "cherry tree" is a flower (or plant) typical for Japanese people, subjects might tend to be quickly reminded of the same words. "Cherry tree" in WordNet is considered as the

lower-level concept of “fruit tree”, and there are about 175 synsets linked to “tree”. Such a categorization yields the increase of the number of links between concepts, hence the distance is found comparatively long. Therefore, the associative concept dictionary, EDR, and WordNet differ from each other in conceptual structure of "cherry tree". When “cherry tree” and “fruit” are excluded, the data cloud of the associative concept dictionary is much closer to WordNet than Goi-Taikei and EDR. (see Fig. 10.)

### 3.7. Comparison using Another Categories

We calculate the distances between two concepts and compare the examples of few categories (“furniture”, ”vegetable”, ”music instrumental” and “bird” in Japanese). Expect “music instrument”, the associative concept dictionary is much closer to WordNet than Goi-Taikei and EDR. The data clouds of four dictionaries about “music instrument” are not closer each ether.

Several observations in the last few sections have shown that the associative concept dictionary and WordNet have closer conceptual structure although there are some differences in culture and in categorization.

## 4. Conclusion

The paper first presents a method to construct an associative concept dictionary using large-scale on-line association experiments. Next the associative concept dictionary is compared with EDR Goi-Taikei and WordNet using the distance between two concepts. The comparison suggests that the associative concept dictionary is more similar to WordNet than Goi-Taikei and EDR. In the Goi-Taikei, many concepts are consolidated at lower-level. Future work includes increasing number of subjects and stimulus words of the association experiment.

## References


[1] Asanoma, N., Alignment of Ontologies: WordNet and Goi-Taikei, NAACL Workshop, 2001.

[2] Budanitsky, A. and Hirst, G., “Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures.”, Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, 2001.

[3] EDR, “EDR Electronic Dictionary Technical Guide”, Japan Electronic Dictionary Research Institute, Ltd., 1990.

[4] S. Ikehara, M. Miyazaki, A. Yokoo, S. Shirai, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi., Nihongo Goi Taikei A Japanese Lexicon. Iwanami Shoten. 5 volumes. (In Japanese), 1997.

[5] George Miller et al., “Five Papers on WordNet”, CSL Report 43, Cognitive Science Laboratory Princeton University, 1993.

[6] Okamoto, J., Ishizaki, S., “Construction of Electronic Concept Dictionary and Quantification of Concept Space”, SIG-NL-130 IPSJ, 1998.

[7] Okamoto, J., Uchiyama, K., Ishizaki, S., “On-line Association Experiment System and Building Concept Dictionary for Basic Vocabulary in Elementary School”, SIG-NL-118 IPSJ, 1997.

[8] Okamoto, J., Ishizaki, S., “Formalization and Modification of Distance Between Two Concepts in Association Concept Space”, IEICE General Conference, D-5-2, IEICE, 2000.

[9] Richardson, R., Alan F. Smeaton and J. Murphy. “Using WordNet for conceptual distance measurement.” Information Retrieval: New Systems and Current Research. Proceedings of the 16th Research Colloquium of the British Computer Information Retrieval Specialist Group, BCSIRSG, pp. 100-123, 1996.

[10] Utiyama, M., Hasida, K., “Bottom-up Alignment of Ontologies”, IJCAI-97 Workshop on Ontologies and Multilingual NLP, 35-40, IJCAI, 1997.