Construction of an Objective Hierarchy of Abstract Concepts via Directional Similarity

Kyoko Kanzaki Eiko Yamamoto Hitoshi Isahara
Computational Linguistics Group,
National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Souraku-gun, Kyoto, Japan,
{kanzaki, eiko, isahara}@nict.go.jp

Qing Ma
Faculty of Science
and Technology
Ryukoku University
Seta, Otsu,520-2194, Japan
qma@math.ryukoku.ac.jp

Abstract

The method of organization of word meanings is a crucial issue with lexical databases. Our purpose in this research is to extract word hierarchies from corpora automatically. Our initial task to this end is to determine adjective hyperonyms. In order to find adjective hyperonyms, we utilize abstract nouns. We constructed linguistic data by extracting semantic relations between abstract nouns and adjectives from corpus data and classifying abstract nouns based on adjective similarity using a self-organizing semantic map, which is a neural network model (Kohonen 1995). In this paper we describe how to hierarchically organize abstract nouns (adjective hyperonyms) in a semantic map mainly using CSM. We compare three hierarchical organizations of abstract nouns, according to CSM, frequency (Tf.CSM) and an alternative similarity measure based on coefficient overlap, to estimate hyperonym relations between words.

1. Introduction

A lexical database is necessary for computers, and even humans, to fully understand a word's meaning because the lexicon is the origin of language understanding and generation. Progress is being made in lexical database research, notably with hierarchical semantic lexical databases such as WordNet, which is used for NLP research worldwide.

When compiling lexical databases, it is important to consider what rules or phenomena should be described as lexical meanings and how these lexical meanings should be formalized and stored electronically. This is a common topic of discussion in computational linguistics, especially in the domain of computational lexical semantics.

The method of organization of word meanings is also a crucial issue with lexical databases. In current lexical databases and/or thesauri, abstract nouns indicating concepts are identified manually and words are classified in a top-down manner based on human intuition. This is a good way to make a lexical database for users with a specific purpose. However, word hierarchies based on human intuition tend to vary greatly depending on the lexicographer, and there is often disagreement as to the make-up of the hierarchy. If we could find an objective method to organize word meanings based on real data, we would avoid this variability.

Our purpose in this research is to extract word hierarchies from corpora automatically. Our initial task to this end is to determine adjective hyperonyms. In order to find adjective hyperonyms, we utilize abstract nouns. Past linguistic research has focused on classifying the semantic relationship between abstract nouns and adjectives (Nemoto 1969, Takahashi 1975).

We constructed linguistic data by extracting semantic relations between abstract nouns and adjectives from corpus data and classifying abstract nouns based on adjective similarity using a self-organizing semantic map (SOM), which is a neural network model (Kohonen 1995). The relative proximity of words in the semantic map indicates their relative similarity.

In previous research, word meanings have been statistically modeled based on syntactic information derived from a corpus. Hindle (1990) used noun-verb syntactic relations, and Hatzivassiloglou and McKeown (1993) used coordinated adjective-adjective modifier pairs. These methods are useful for the organization of words deep within a hierarchy, but do not seem to provide a solution for the top levels of the hierarchy.

To find an objective hierarchical word structure, we utilize the complementary similarity measure (CSM), which estimates a one-to-many relation, such as superordinate–subordinate relations (Hagita and Sawaki 1995, Yamamoto and Umemura 2002).

In this paper we propose an automated method for constructing adjective hierarchies by connecting strongly related abstract nouns in a top-down fashion within a semantic map, mainly using CSM. We compare three hierarchical organizations of abstract nouns, according to CSM, frequency (Tf.CSM) and an alternative similarity measure based on coefficient overlap, to estimate hyperonym relations between words.

2. Linguistic clues to extract adjective hyperonyms from corpora

In order to automatically extract adjective hyperonyms we use syntactic and semantic relations between words.

There is a good deal of linguistic research focused on the syntactic and semantic functions of abstract nouns, including Nemoto (1969), Takahashi (1975), and Schmid (2000). Takahashi (1975) illustrated the sentential function of abstract nouns with the following examples.

a. *Yagi wa seishitsu ga otonashii.*
(goat) topic (nature) subject (gentle)
The nature of goats is gentle

b. *Zou wa hana ga nagai.*
(elephant) topic (a nose) subject (long)
The nose of an elephant is long

He examined the differences in semantic function between "*seishitsu* (nature)" in (a) and "*hana* (nose)" in (b), and explained that "*seishitsu* (nature)" in (a) indicates an aspect of something, i.e., the goat, and "*hana* (nose)" in (b) indicates part of something, i.e., the elephant. He recognized abstract nouns in (a) as a hyperonym of the attribute that the predicative adjectives express. Nemoto (1969) identified expressions such as "*iro ga akai* (the color is red)" and "*hayasa ga hayai* (the speed is fast)" as a kind of meaning repetition, or tautology.

In this paper we define such abstract nouns that co-occur with adjectives as adjective hyperonyms. We semi-automatically extracted from corpora 365 abstract nouns used as this kind of head noun, according to the procedures described in Kanzaki et al. (2000). We collected abstract nouns from two year's worth of articles from the Mainichi Shinbun newspaper, and extracted adjectives co-occurring with abstract nouns in the manner of (a) above from 100 novels, 100 essays and 42 year's worth of newspaper articles, including 11 year's worth of Mainichi Shinbun articles, 10 year's worth of Nihon Keizai Shinbun (Japanese economic newspaper) articles, 7 year's worth of Sangyoukinyuuryuutsu Shinbun (an economic newspaper) articles, and 14 year's worth of Yomiuri Shinbun articles. The total number of abstract noun types is 365, the number of adjective types is 10,525, and the total number of adjective tokens is 35,173. The maximum number of co-occurring adjectives for a given abstract noun is 1,594.

3. On the Self-Organizing Semantic Map

3.1 Input data

Abstract nouns are located in the semantic map based on the similarity of co-occurring adjectives after iteratively learning over input data.

In this research, we focus on abstract nouns co-occurring with adjectives. In the semantic map, there are 365 abstract nouns co-occurring with adjectives. The similarities between the 365 abstract nouns are determined according to the number of common co-occurring adjectives. We made a list such as the following.

OMOI (feeling): *ureshii* (glad), *kanashii* (sad), *shiawasena* (happy), …

KIMOCHI (though): *ureshii* (glad), *tanoshii* (pleased), *hokorashii* (proud), …

KANTEN (viewpoint): *igakutekina* (medical), *rekishitekina* (historical), ...

When two (or more) sets of adjectives with completely different characteristics co-occur with an abstract noun and the meanings of the abstract noun can be distinguished correspondingly, we treat them as two different abstract nouns. For example, the Japanese abstract noun "men" is treated as two different abstract nouns with "men1" meaning "one side (of the characteristics of someone or something)" and "men2" meaning "surface". The former co-occurs with "gentle", "kind" and so on. The latter co-occurs with "rough", "smooth" and so on.

3.2 The Self-Organizing Semantic Map

Ma (2000) classified co-occurring words using a self-organizing semantic map (SOM).

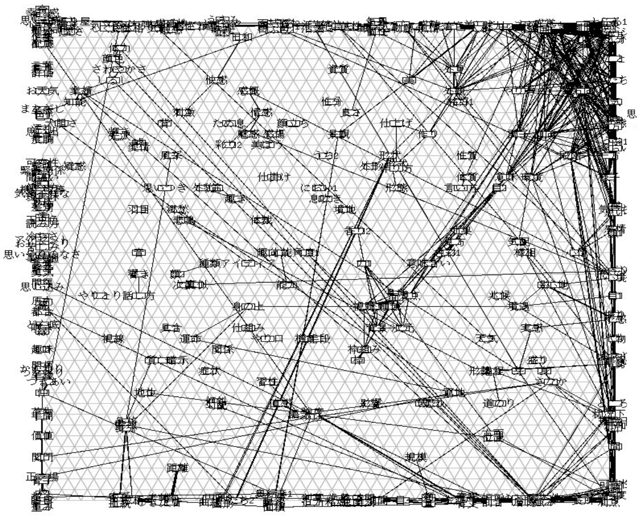

Figure 1. The Cosine-based SOM of word similarity

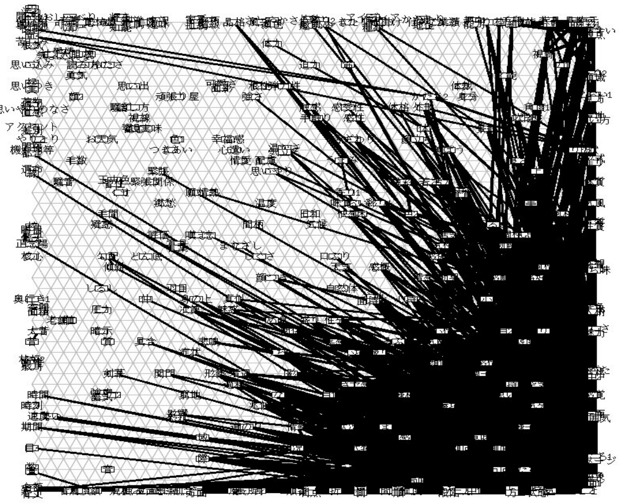

Figure 2. The CSM-based SOM of word similarity

We made a semantic map of the above-mentioned 365 abstract nouns using SOM, based on the cosine measure. The distribution of the words in the map gives us a sense of the semantic distribution of the words. However, we could not precisely identify the relations between words in the map (Fig 1). In Fig. 1 lines on the maps indicate close relations between word pairs. In the cosine-based semantic map, there is no clear correspondence between word similarities and the distribution of abstract nouns in the map.

To solve this problem we introduced the complementary similarity measure (CSM). This similarity measure estimates one-to-many relations, such as superordinate–subordinate relations (Hagita and Sawaki 1995, Yamamoto and Umemura 2002). We can find the hierarchical distribution of words in the semantic map according to the value of CSM (Fig 2). In the CSM-based SOM, lines are concentrated at the bottom right hand corner, that is, most abstract nouns are located at the bottom right-hand corner.

Next, we find hierarchical relations between whole abstract nouns, not between word pairs, on the map automatically.

4. How to construct hierarchies of nominal adjective hyperonyms in the Semantic Map

4.1 Similarity measures, CSM and Yates' correction

A feature of CSM is its ability to estimate hierarchical relations between words. This similarity measure was developed for the recognition of degraded machine-printed text (Hagita and Sawaki, 1995). Yates' correction is often used in order to increase the accuracy of approximation. Hierarchical relations can be extracted accurately when the CSM value is high. Yates' correction can extract different relations from high CSM values. When the CSM value is low, the result is not reliable, in which case we use Yates' correction.

According to Yamamoto and Umemura (2002), who adopted CSM to classify words, CSM is calculated as follows.

$$CSM = \frac{ad - bc}{\sqrt{(a+c)(b+d)}}$$

Yates' correction is calculated as follows.

$$Yates = \frac{n(|ad - bc| - n/2)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Here n is the sum of the number of co-occurring adjectives; a indicates the number of times the two labels appear together; b indicates the number of times "label 1" occurs but "label 2" does not; c is the number of times "label 2" occurs but "label 1" does not; and d is the number of times neither label occurs. In our research, each "label" is an abstract noun, a indicates the number of adjectives co-occurring with both abstract nouns, b and c indicate the number of adjectives co-occurring with either abstract noun ("label 1" and "label 2", respectively), and *d* indicates the number of adjectives co-occurring with neither abstract noun. We calculated hierarchical relations between word pairs using these similarity measures.

4.2 Construction of a hierarchy of abstract nouns using CSM and Yates' correction

The hierarchy construction process is as follows:

1) Based on the results of CSM, "*koto* (matter)" is the hyperonym of all abstract nouns.

First, we connect super/sub-ordinate words with the highest CSM value while keeping the super-subordinate relation.

2) When the normalized value of CSM is lower, the number of extracted word pairs becomes increasing overwhelmingly, and the reliability of CSM diminishes. Word pairs with a normalized CSM value of less than 0.4 are located far from the common hyperonym "*koto* (matter)" on the semantic map. If we construct a hierarchy using CSM values only, a long hierarchy containing irrelevant words emerges. In this case, the word pairs calculated by Yates' correction are more accurate than those from CSM. We combine words using Yates' correction, when the value of CSM is less than 0.4. When we connect word pairs with a high Yates' value, we find a hyperonym of the super-ordinate noun of the pair and connect the pair to the hyperonym. If a word pair appears only in the Yates' correction data, that is, we cannot connect the pair with high Yates' value to the hyperonym with high CSM value, they are combined with "*koto* (matter)".

3) Finally, if a short hierarchy is contained in a longer hierarchy, it is merged with the longer hierarchy and we insert "*koto* (matter)" at the root of all hierarchies.

4.3 Results

The number of groups obtained was 161. At its deepest, the hierarchy was 15 words deep, and at its shallowest, it was 4 words deep. The following is a breakdown of the number of groups at different depths in the hierarchy.

The greatest concentration of groups is at depth 7. There are 140 groups from depth 5 to depth 10, which is 87% of all groups.

Table 1: The depth of the hierarchy by CSM

Depth	***4***	***5***	***6***	*7*	***8***	***9***
Groups	3	16	27	32	23	23
Depth	***10***	***11***	***12***	***13***	***14***	***15***
Groups	19	7	3	4	3	1

The word that has the strongest relation with "*koto* (matter)" is "*men1* (side1)". The number of groups in which "*koto* (matter)" and "*men1* (side1)" are hyperonyms is 96 (59.6%). The largest number of groups after that is a group in which "*koto* (matter)", "*men1* (side1)" and "*imeeji* (image)" are hyperonyms. The number of groups in this case is 59 groups, or 36.6% of the total. With respect to the value of CSM, the co-occurring adjectives are similar to "*men1* (side1)" and "*imeeji* (image)".

Other words that have a direct relation with "*koto* (matter)" are "*joutai* (state)" and "*toki* (when)". They have the most number of groups after "*men1* (side1)" among all the children of "*koto* (matter)". The number of groups subsumed by "*joutai* (state)" group and "*toki* (when)" are 21 and 19, respectively. Other direct hyponyms of "*koto* (matter)" are:

ki (feeling): 6 groups
ippou (while or grow –er and er): 3 groups
me2 (eyes): 3 groups
katachi1 (in the form of): 3 groups
iikata (how to say): 2 groups
yarikata (how to): 2 groups

There is little hierarchical structure to these groups, as they co-occur with few adjectives.

4.4 The Hierarchies of abstract concepts in the semantic map

In the following semantic maps, where abstract nouns are distributed using SOM and CSM (see Section 3), hierarchies of abstract nouns are drawn with lines. The bottom right hand corner is "*koto* (matter)", a starting point for the distribution of abstract nouns.

Five main types of hierarchies are found from patterns of lines on the map, as follows:

The first figure, Fig.3, is hierarchies of "*kanji* (feeling), *kimochi* (feeling) …" on the semantic map. The location of hierarchies of "*yousu* (aspect), *omomochi* (look*), kaotsuki* (on one's face), …" is similar to this type of the location. Hierarchies of "*sokumen* (one side), *imi* (meaning), *kanten* (viewpoint), *kenchi* (standpoint) …" on

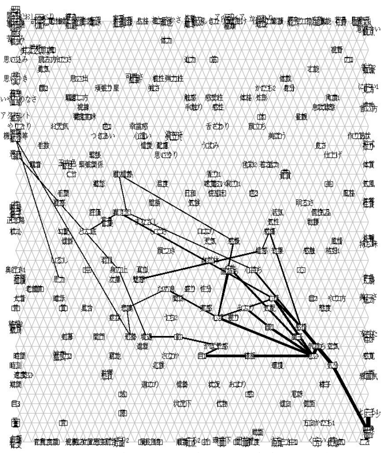

Fig.3: Hierarchies of "*kimochi* (feeling)"

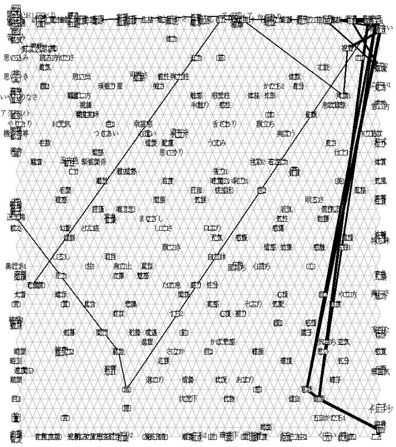

Fig.4:Hierarchies of "*sokumen* (one side)"

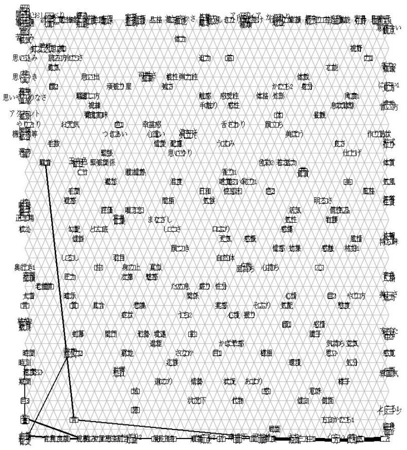

Fig.5:Hierarchies of "*teido* (degree)"

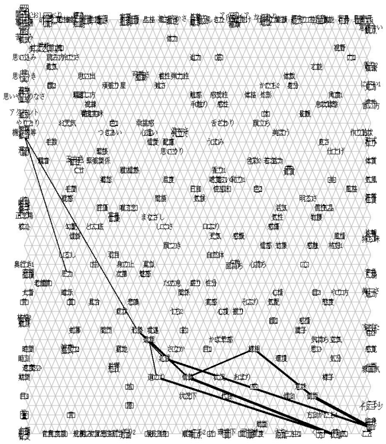

Fig.6: Hierarchies of "*jousei* (situation)"

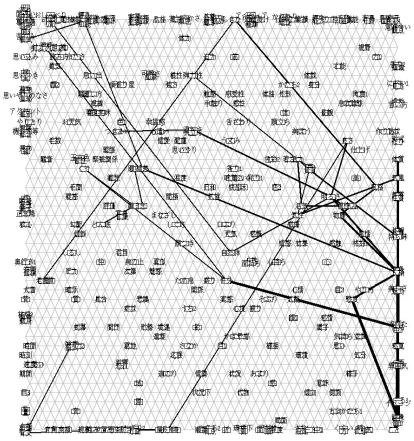

Fig.7:Hierarchies of "*seikaku* (character)"

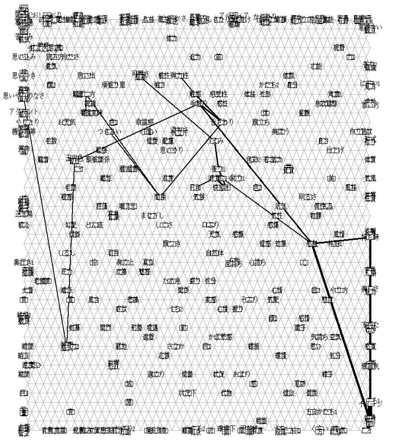

Fig8: Hierarchies of "*kanshoku* (feel)"

the map are shown in Fig. 4. The lines of the hierarchies go up from the bottom right hand corner to the upper left hand corner and then turn towards the upper right hand corner. The location of hierarchies of "*nouryoku* (ability), *sainou* (talent) …" is similar to this one.

The hyperonym of "*teido* (degree)" is "*joutai* (state)". In Fig.5 these abstract nouns are located at the bottom of the map. The location of hierarchies of "*kurai* (rather than)" and "*hou* (comparatively)" are similar to this one. The hierarchies of "*joutai* (state), *joukyou* (situation), *yousou* (aspect), *jousei* (the state of affairs)" are shown in Fig.6. The lines are found at a higher location than the line of "*teido*(degree)". The lines of the hierarchies of "*joutai* (state), *ori* (when), *sakari* (in the hight of), *sanaka* (while)" are similar to these lines.

The lines of the hierarchies of "*seikaku* (character)", "*gaikan* (appearance)"and "*utsukushisa* (beauty)" are similar to each other. We show the hierarchies of "*seikaku* (character)" in Fig.7. These lines in Fig.7 are located from the right end to the upper left hand corner. From the following, we can find five main types of hierarchies.

From the starting point " *koto* (matter)",

-The hierarchies of "*men* (side), *inshou* (impression), *kanji* (feeling), *kibun* (mood), *kimochi* (feeling)"
-The hierarchies of "*men* (side), *sokumen* (one-side), *imi* (meaning), *kanten* (viewpoint), *kenchi* (standpoint)"
-The hierarchies of "*joutai* (state), *teido* (degree)"
-The hierarchies of "*joutai* (state), *jousei* (situation)"
-The hierarchies of "*men* (side), *inshou* (impression), *seikaku* (character) or *gaikan* (appearance) or *utsukushisa* (beauty)".

The lines in Fig.8 are not peculiar, and appear in an area of the hierarchies of "*seikaku* (character)” in Fig.7. As Fig.8 shows, the hierarchies of “*men* (side), *inshou* (impression), *kanji* (feeling), *kanshoku* (feel) or *kansei* (sensitivity)” are located in the area of the hierarchies of “*seikaku* (character)”, above the hierarchies of “*kimochi* (feeling)” in Fig.3.

5. Comparison of hierarchies of superordinate nouns of adjectives.

We compare the hierarchy mentioned above with ones obtained from two kinds of data.

1) Hierarchies obtained by:
CSM and Yate’s correction
corpus occurrence data (no frequency).
2) Hierarchies obtained by:
Tf.CSM and Yate’s correction
corpus frequency data.
3) Hierarchies obtained by:
Overlap coefficient and Yates' correction
corpus occurrence data (no frequency).

As both CSM and the Overlap coefficient are “measures of inclusion”, we compared CSM and Tf.CSM with the Overlap coefficient.

The number of groups that were obtained by CSM, Tf.CSM and the Overlap coefficient are the following.

Table 2. Total number of groups obtained from CSM, Tf.CSM and Ovlp (Overlap)

	groups
CSM	161
Tf.CSM	158
Ovlp	240

The Depth of hierarchies obtained from CSM, Tf.CSM, and the Overlap coefficient are as follows:

Table 3. The hierarchy depth for CSM, Tf.CSM, and the Overlap coefficient

depth	***3***	***4***	***5***	***6***	***7***	***8***	***9***
CSM	0	3	16	27	32	23	23
Tf.CSM	1	5	10	18	13	25	11
Ovlp	32	56	61	57	21	7	2

depth	***10***	***11***	***12***	***13***	***14***	***15***
CSM	19	7	3	4	3	1
Tf.CSM	24	13	14	14	7	2
Ovlp	2	0	0	0	0	0

In the case of CSM, there are 32 groups at depth 7, which is the greatest number of groups. The greatest concentration of groups is at depth 5 to 10. In the case of Tf.CSM, the greatest number of groups is 25 at depth 8. The greatest concentration of groups is at depth 5 to 13. In the case of the overlap coefficient, the greatest number of groups is 61 at depth 5. The greatest concentration of groups is at depth 3 to 7.

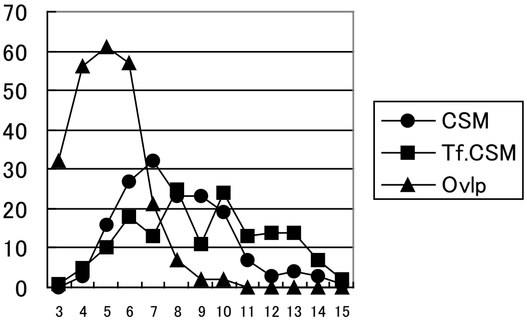

Figure 9. Distribution of hierarchy depth for CSM, Tf.CSM, and Overlap coefficient

From this result, we can see that hierarchies generated by Tf.CSM are relatively deep, and those generated by the Overlap coefficient are relatively shallow.

In the case of the Overlap coefficient, abstract nouns in lower layers are sometimes directly related to abstract nouns in the highest layers. On the other hand, in hierarchies generated by CSM and Tf.CSM, abstract nouns in the highest layers are related to those in the lowest layers via abstract nouns in the middle layers. The following indicates the number of overlapping hierarchies for CSM, Tf.CSM and Overlap.

Table 4. The number of overlapping hierarchies among CSM, Tf.CSM and Overlap

CSM&Tf.CSM	37
CSM&Ovlp	7
Tf.CSM&Ovlp	2
CSM&Tf.CSM&Ovlp	7

The hierarchy generated by Tf.CSM is the deepest, and includes some hierarchies generated by CSM and the Overlap coefficient. The hierarchy generated by CSM is more similar to the one made by Tf.CSM than that for the Overlap coefficient: the number of completely corresponding hierarchies for CSM and Tf.CSM is 37, that for CSM and the Overlap coefficient is 7, and that for Tf.CSM and the Overlap coefficient is 2. The total number of hierarchies that correspond completely between CSM, Tf.CSM and the Overlap coefficient is 7, and the number of hierarchies which are generated by two of the methods and included in the third is 57.

We investigated these 64 hierarchies precisely, checking adjectives appearing at each depth as indicated by an abstract noun in this paper. In 6 of these hierarchies, the same adjectives were found at all levels of the hierarchy. In 14 of the remaining 58 hierarchies, the same adjectives were found in all but the deepest level. These 20 hierarchies are the most plausible in the strict sense of the word. Below, we give examples of these hierarchies. In the next stage of this research, we intend to investigate the remaining 44 hierarchies to determine the reason for the difference in adjective content.

The common hyperonym: *koto* (matter) ---
*men*1 (side) ---
sokumen (one side) ---
imi (meaning) ---
kanten (viewpoint) ---
*me*2 (eyes) ---
mikata (view) ---
hyouka (evaluation) ---
ippou (while or grow -er and er) ---
ikioi (force) ---
sokudo (speed) ---
jikoku (time) ---

6. Conclusion

We have suggested how to make a hierarchy of adjectives automatically by connecting strongly-related abstract nouns in a top-down fashion. We generated a word hierarchy from corpus data by using a combination of two methods: a self-organizing semantic map and a directional similarity measure. As our directional similarity measure, we utilized the complementary similarity measure (CSM). Then we compared the hierarchy generated by CSM with that generated by Tf.CSM and the Overlap coefficient. In the case of Tf.CSM, the hierarchy is deeper than the others because there are more abstract nouns in the middle layer. In the case of the Overlap coefficient, the hierarchy is shallow, but there are more hyponyms in the lower layer than with the other two methods. As a result, the hierarchies generated by CSM have more common hierarchical relations than those generated by the other two methods. In future work, we will analyze common hierarchies made by the three methods in detail and examine differences among them in order to generate an abstract conceptual hierarchy of adjectives. We will then compare our hierarchy with thesauri compiled manually. After we have completed the experiment on Japanese adjectives, we are keen to investigate differences and similarities in adjective hyperonyms between Japanese and other languages such as English by means of our method.

Acknowledgement

We would like to thank Dr. Masaki Murata of NICT for allowing us to use his drawing tool.

References

Nemoto, K. 1969. The combination of the noun with “ga-Case” and the adjective, *Language research2 for the computer*, National Language Research Institute: 63-73/

Takahashi, T. 1975. A various phase related to the part-whole relation investigated in the sentence, *Studies in the Japanese language* 103, The society of Japanese Linguistics: 1-16.

Kohonen, T. 1995. *Self-Organizing Maps*, Springer.

Hindle, D. 1990. Noun Classification From Predicate-Argument Structures, *In the Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*: 268-275

Hatzivassiloglou,V. and McKeown,R.K. 1993. Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning, *In the Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*: 172-182.

Hagita, N. and Sawaki, M. 1995. Robust Recognition of Degraded Machine-Printed Characters using Complimentary Similarity Measure and Error-Correction Learning, *In the Proceedings of the SPIE –The International Society for Optical Engineering*, 2442: 236-244.

Yamamoto, E. and Umemura, K. 2002. A Similarity Measure for estimation of One–to-Many Relationship in Corpus, *Journal of Natural Language Processing*: 45-75.

Hans-Jorg Shmid. 2000. *English Abstract Nouns as Conceptual Shells*, Mouton de Gruyter.

Kanzaki, K., Ma., Q. and Isahara, H. (2000), Similarities and Differences among Semantic Behaviors of Japanese Adnominal Constituents, *In the Proceedings of the Syntactic and Semantic Complexity in Natural Language Processing Systems, ANLP and NAACL.*

Ma, Q., Kanzaki, K., Murata, M., Uchimoto, K. and Isahara, H. 2000. Self-Organization Semantic Maps of Japanese Noun in Terms of Adnominal Constituents, *In Proceedings of IJCNN’2000*, Como, Italy, vol.6.: 91-96.