Extraction of hyperonymy of Adjectives from Large Corpora by Using the Neural Network Model

Kyoko Kanzaki[*]**, Qing Ma**[†]**, Eiko Yamamoto**[*]**, Masaki Murata**[*] **and Hitoshi Isahara**[*]

[*] **National Institute of Information and Communications Technology**
3-5, Hikaridai, Seikacho, Sourakugun, Kyoto, 619-0289, Japan
{kanzaki, eiko, murata, isahara} @crl.go.jp

[†] **Ryukoku University**
Seta, Otsu,520-2194, Japan
qma@math.ryukoku.ac.jp

Abstract

In this research, we extract hierarchical abstract concepts of adjectives automatically from large corpora by using the Neural Network Model. We show the hierarchies on the Semantic Map and compare the hierarchies in the Semantic Map and a manually prepared thesaurus. We recognized five types of distributions on the map. By comparing the Semantic Map and a manual thesaurus, we found that the word that the abstract noun belongs to, whether a person, thing or event, is introduced as the standard of classification in the manual thesaurus. On the other hand, in the Semantic Map, we found that abstract nouns belonging to people or events are distributed together. We also found that the hierarchies of *sokumen* (side), *imi* (meaning), and *kanten* (viewpoint) are necessary for a category of adjectives.

1. Introduction

Constructing a thesaurus automatically and accurately from real data is a crucial issue in compiling a lexical database. For our purposes, we need to not only find the similarity between words but also the hierarchical relationship between words. In this research, we extract hierarchical abstract adjective concepts automatically from large corpora using the Neural Network Model (Kohonen 1995, Ma 2000). We define abstract nouns co-occurring with adjectives as our determinant of abstract adjective semantics. We show the hierarchies of abstract adjective semantics within the Semantic Map, and compare hierarchies in the Semantic map with those in a manually prepared thesaurus.

2. Abstract Concepts of adjectives

In order to automatically extract adjective hypernyms, 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 hypernym 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 hypernyms. We extracted these co-occurrence relations between abstract nouns and adjectives from corpus data automatically, and used them as input data to SOM.

We extracted 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 Sangyoukinyuuryuutusu 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. 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), ...

We use this linguistic data as the input data to SOM.

3.2. The Semantic Map by using CSM

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

We made a semantic map of the above-mentioned 365 abstract nouns using SOM. However, we could not precisely identify the relations between words in the map. The distribution of the words on the map gives us a clue as to the semantic distribution of the words.

To solve this problem, we introduced the complementary similarity measure (CSM). This similarity measure estimates a one-to-many relation, 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.

This similarity measure was developed for the recognition of degraded machine-printed text (Hagita and Sawaki, 1995). Hierarchical relations can be extracted accurately when the CSM value is high. When the CSM value is low, however, the result is unreliable. To compensate for this weakness of CSM, we use Yates' correction. Yates' correction is often used in order to increase the accuracy of approximation. Yates' correction can extract different relations from high CSM values.

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. How to obtain hierarchical relations

The hierarchy construction process is as follows:

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

We therefore choose words which have a high CSM value with "*koto* (matter)" and connect them to it. Here, "*koto* (matter)" is a hyperonym of these words. Next, we choose words which have a high CSM value with these new words, and so on iteratively.

2) When the normalized value of CSM is less than 0.4, the number of extracted word pairs becomes increasingly overwhelming, and the reliability of CSM diminishes. Word pairs with a normalized CSM value of less than 0.4 are located far from the common hypernym "*koto* (matter)" on the semantic map. If we construct a hierarchy using CSM value 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 the hyperonym of the superordinate noun in 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 a high Yates' value to the hyperonym with a 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.

5. A Hierarchy of abstract concepts of adjcetives

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.

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

Table 1: The depth of the hierarchy obtained by CSM

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.

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 hypernyms is 96 (59.6%). The largest number of groups after that is a group in which "*koto* (matter)", "*men1* (side1)" and "*imeeji* (image)" are hypernyms. 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 (have a experience)	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.

6. The Hierarchies of abstract concepts on the Semantic Map

The Semantic Map can be divided into three basic regions, based on the distribution of abstract nouns.
The bottom right hand corner is "*koto* (matter)", a starting point for the distribution of abstract nouns. In the middle of the map, abstract nouns relating to a mental state are distributed. In the upper right-hand area, abstract nouns concerning an impression of someone or something are located. In the area in the bottom left-hand corner, abstract nouns about state are located.

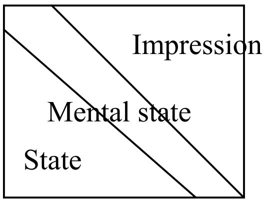


Fig1: The rough sketch of semantic areas

On the following semantic maps, hierarchies of abstract nouns are drawn with lines.

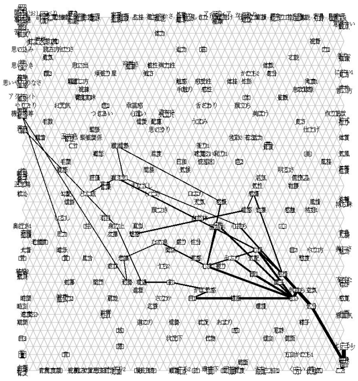
Five main types of hierarchies are found, as follows:

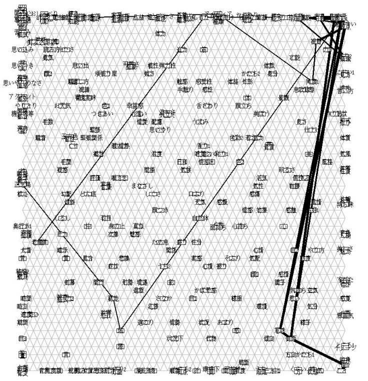
The first figure, Fig.2, is hierarchies of "*kanji* (feeling), *kimochi* (feeling) …" on the Semantic Map. The location of hierarchies of "*yousu* (aspect), *omomochi* (look*), kaotsuki* (on one's face), …" is similar to this type of the location.

Hierarchies of "*sokumen* (one side), *imi* (meaning), *kanten* (viewpoint), *kenchi* (standpoint) …" on the map are shown in Fig. 3. 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.4 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.5. The lines are found at a higher location than the line of "*teido*(degree)". The lines of hierarchies of "*joutai* (state), *ori* (when), *sakari* (in the hight of), *sanaka* (while)" are similar to these lines.





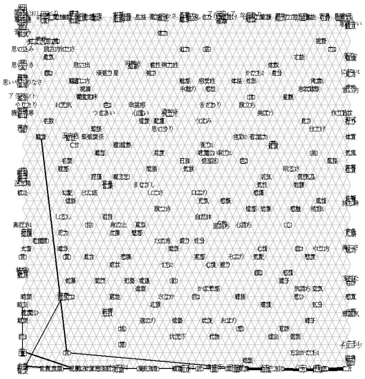
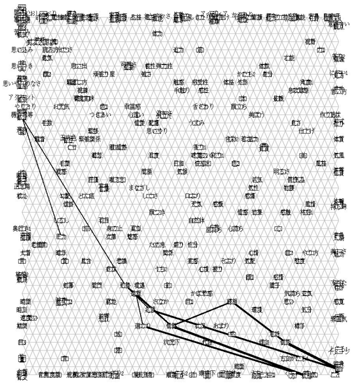
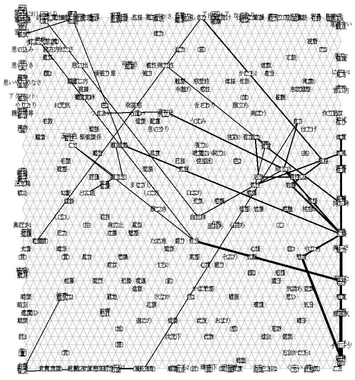


Fig.2: Fig.3: Fig.4:

Hierarchies of "*kimochi* (feeling)" Hierarchies of "*sokumen* (one side)" Hierarchies of "t*eido* (degree)"





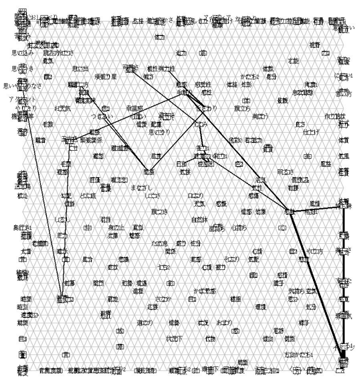


Fig.5: Fig.6: Fig.7:

Hierarchies of "*jousei* (situation)" Hierarchies of "*seikaku* (character)" Hierarchies of "*kanshoku* (feel)"

The lines of hierarchies of "*seikaku* (character)", "*gaikan* (appearance)"and "*utsukushisa* (beauty)" are similar to each other. We show the hierarchies of "*seikaku* (character)" in Fig.6. These lines in Fig6 are located from the right end to the upper left hand corner.

In Fig.7 the hierarchies of "*kanshoku* (feel)" are shown. As the lines start from the right end, they are similar to hierarchies of "seikaku(character)". They are then located higher than hierarchies of a mental state in Fig.2.From the above, 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 hierarchies of "*men* (side), *inshou* (impression), *kanji* (feeling), *kanshoku* (feel) or *kansei* (sensitivity)" are located near both the hierarchies of "*seikaku* (character)" and "*kimochi* (feeling)".

7. The comparison between thesaurus by SOM + CSM and by human

We compare the semantic map with two manual thesauri, i.e., BUNRUIGOIHYOU (WORD LIST BY SEMANTIC PRINCIPLE) and the EDR lexicon.

In BUNRUIGOIHYOU, adjectives and adverbs are listed together. They belong to groups under the category SOU (modifier). They are divided into three main categories:

CHUSHOU_TEKI KANKEI (an abstract relation)
Relationship, pace, goodness/badness, evaluation of qualification, time, location, shape, degree et al.

SEISHIN OYOBI KOUI (mentality/action)
Personality, attitude, mental state, sense, evaluation of human action, circumstances, association et al.

SHIZENGENSHOU (nature)
Light, color, taste, smell, sound, material, body et al.

These categories, i.e. abstract adjective concepts, are determined based on whether the abstract nouns refer to a person or not.

The EDR lexicon is one of the largest Japanese electronic lexicons. This lexicon has a manual thesaurus. It has hierarchical relations for "concept", "concepts related to information processing", "phenomena", "matter", "time", "position", "an agent such as a person and a creature behaving like a person", and "attributes, values and units in the domain of information processing" and their subordinate concepts. They are divided into hyponymic concepts based on whether a reference of an abstract noun belongs to a person or to a thing.

With a Semantic Map, the distribution of abstract nouns on the map is not based on what the reference of an abstract noun belongs to. For example, as "*takai* (tall/high)" co-occurrs with "*setake* (the height of human)" and with "*kingaku* (an amount of money)", they are separated in the manual thesaurus but they are closely located in the automatically-constructed Semantic Map. They are included in the hierarchy of "*joutai* (situation) and *teido* (degree)".

In our data, from the root node "*koto* (matter)", abstract nouns are divided into five hierarchies; mental expressions, impression of something or someone, states similar to situations, states like degrees, and a meaning/viewpoint.

8. Conclusion

In this research, we distributed abstract concepts on a Semantic Map by using CSM, and derived a hierarchy of abstract concepts by using CSM and Yates. On the map we recognized five types of distributions. When we compared the Semantic Map with a manual thesaurus, we found that what the type of the abstract noun referent, i.e., a person, thing or event, is used as the primary classification criterion in manual thesauri. On the other hand, in the Semantic Map, we found that abstract nouns that refer to both people and events are distributed together. A criterion for the classification of abstract concepts is a mental state or impression of something, someone or a state, or a viewpoint of something. When considering adjectives, these seem to be used flexibly without regard to a person or event.

The hierarchy of "*sokumen* (a side), *imi* (meaning) and *kanten* (viewpoint)" is necessary for categorizing adjectives, because co-occurring adjectives are a peculiarity. For example, "*rekishitekin* (historical)", "*igakutekina* (medical)", "*seijitekina* (political)" and so on occur. In future work, we will attempt to compare hierarchies by using CSM with other methods.

References

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

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.

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.

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

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.

National Language Research Institue. 1964. *WORD LIST BY SEMANTIC PRINCIPLE*, Shuei Shuppan.

EDR electronic Dictionary.
URL: http://www2.crl.go.jp/kk/e416/EDR/J_index.html