Extraction of Hierarchies Based on Inclusion of Co-occurring Words with Frequency Information

Eiko Yamamoto Kyoko Kanzaki Hitoshi Isahara
Computational Linguistics Group
National Institute of Information and Communications Technology,
3-5 Hikari-dai, Seika-cho, Souraku-gun, Kyoto 619-0289, Japan
eiko@nict.go.jp kanzaki@nict.go.jp isahara@nict.go.jp

Abstract

In this paper, we propose a method of automatically extracting word hierarchies based on the inclusion relations of word appearance patterns in corpora. We applied the complementary similarity measure (CSM) to determine a hierarchical structure of word meanings. The CSM is a similarity measure developed for recognizing degraded machine-printed text. There are CSMs for both binary and gray-scale images. The CSM for binary images has been applied to estimate one-to-many relations, such as superordinate-subordinate relations, and to extract word hierarchies. However, the CSM for gray-scale images has not been applied to natural language processing. Here, we apply the latter to extract word hierarchies from corpora. To do this, we used frequency information for co-occurring words, which is not considered when using the CSM for binary images. We compared our hierarchies with those obtained using the CSM for binary images, and evaluated them by measuring their degree of agreement with the EDR electronic dictionary.

1 Introduction

The hierarchical relations of words are useful language resources. Progress is being made in lexical database research, notably with hierarchical semantic lexical databases such as WordNet [Miller *et al.*, 1990] and the EDR electronic dictionary [1995], which are used for natural language processing (NLP) research worldwide. These databases are essential for enabling computers, and even humans, to fully understand the meanings of words because the lexicon is the origin of language understanding and generation. In current thesauri with hierarchical relations, words are categorized manually and classified in a top-down manner based on human intuition. This may be a practical way of developing a lexical database for NLP. However, these word hierarchies tend to vary greatly depending on the lexicographer. In fact, each thesaurus includes original hierarchical relations that differ from those in other thesauri. There is often disagreement as to the make-up of a hierarchy. In addition, hierarchical relations based on different data may be needed depending on the user. A statistical method of creating hierarchies from corpora would thus be useful. We therefore attempted to automatically extract hierarchies most suited to the information that a user handles. To do this, we extract hypernym-hyponym relations between two words from corpora and then build hierarchies by connecting these relations. As the initial task, we attempted to extract hierarchies of abstract nouns that co-occur with adjectives in Japanese.

In finding word hierarchies in corpora, it is usual to use patterns, such as "a part of," "is a," "such as," or "and," obtained from the corpora [Hearst, 1992; Berland and Charniak, 1999; Caraballo, 1999]. Methods for extracting hypernyms of entry words from definition sentences in dictionaries [Tsurumaru *et al.*, 1986; Shoutsu *et al.*, 2003] and methods using collocations retrieved from corpora [Nakayama and Matsumoto, 1997] have been described previously. A hybrid method that uses both dictionaries and the dependency relations of words taken from a corpus has also been reported [Matsumoto *et al.*, 1996]. Recently, a similarity measure developed for recognizing degraded machine-printed text [Hagita and Sawaki, 1995] was used to estimate one-to-many relations, such as that of superordinate-subordinate, from a corpus [Yamamoto and Umemura, 2002]. This measure is called the complementary similarity measure (CSM) for binary images and indicates the degree of the inclusion relation between two binary vectors. In that study, each binary vector corresponds to certain appearance patterns for each term in a corpus. The CSM for binary images has also been applied to extract word hierarchies from corpora [Yamamoto *et al.*, 2004] and to trace the distribution of abstract nouns on a self-organizing semantic map [Kanzaki *et al.*, 2004].

In the experiments described in this paper, we attempted to use the CSM for gray-scale images [Sawaki *et al.*, 1997] to extract hypernym-hyponym relations between two words. Specifically, we used not only binary vectors with elements of 0 or 1, but also vectors consisting of weights based on the frequencies of co-occurring words in corpora. We compared the hierarchies extracted using the CSM for gray-scale images with those extracted using the CSM for binary images. Finally, to verify the effectiveness of our approach, we

evaluated our hierarchies by measuring the degree to which they agreed with the EDR electronic dictionary.

2 Experimental Data

A good deal of linguistic research has focused on the syntactic and semantic functions of abstract nouns [Nemoto, 1969; Takahashi, 1975; Kanzaki *et al.*, 2003]. In the example, "*Yagi* (goat) *wa seishitsu* (nature) *ga otonashii* (gentle) (Goats have a gentle nature)," Takahashi [1975] recognized that the abstract noun "*seishitsu* (nature)" is a hypernym of the attribute expressed by the predicative adjective "*otonashi* (gentle)." To classify adjectives on the basis on these functions, Kanzaki *et al.* [2003] defined such abstract nouns that co-occur with adjectives as hypernyms of these adjectives. They produced linguistic data for their research by automatically extracting the co-occurrence relations between abstract nouns and adjectives from corpora.

In our experiment, we used the same corpora and abstract nouns as Kanzaki *et al.*. The corpora consist of 100 novels, 100 essays, and 42 years' worth of newspaper articles, including 11 years of the Mainichi Shinbun, 10 years of the Nihon Keizai Shinbun, 7 years of the Sangyou-kinyuuryuutsuu Shinbun, and 14 years of the Yomiuri Shinbun. The abstract nouns were selected from 2 years' worth of the Mainichi Shinbun newspaper articles by Kanzaki *et al.*.

However, we produced our experimental data using only the sentences in the corpora that could be parsed using KNP, which is a Japanese parser developed at Kyoto University. The parsed data consisted of 354 abstract nouns and 6407 adjectives and included the following examples (the number after each adjective is the frequency which the abstract noun co-occurs with the adjective):

OMOI (feeling): *ureshii* (glad) 25, *kanashii* (sad) 396, *shiawasena* (happy) 6, …

KIMOCHI (thought): *ureshii* (glad) 204, *tanoshii* (pleasant) 87, *hokorashii* (proud) 40, …

KANTEN (viewpoint): *igakutekina* (medical) 9, *rekishitekina* (historical) 17, ...

3 Complementary Similarity Measure

As mentioned above, we used the complementary similarity measure (CSM) to estimate the hierarchical relations between word pairs. The CSM was developed for recognizing degraded machine-printed text [Hagita and Sawaki, 1995; Sawaki *et al.*, 1997]. There are two kinds of CSMs, one for binary images and one for gray-scale images.

3.1 Complementary similarity measure for binary images

The CSM for binary images was developed as a character recognition measure for binary images and is designed to be robust against heavy noise or graphical design [Hagita and Sawaki, 1995]. It was applied to estimate one-to-many relationships between words by Yamamoto and Umemura [2002]. They estimated these relations from the inclusion relations between the appearance patterns of two words. An appearance pattern is expressed as an n-dimensional binary feature vector. Let $F = (f_1, f_2, \ldots, f_i, \ldots, f_n)$, where $f_i = 0$ or 1, and let $T = (t_1, t_2, \ldots, t_i, \ldots, t_n)$, where $t_i = 0$ or 1, be the feature vectors of the appearance patterns for two words. The *CSM* of F to T is then defined as follows:

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

$$a = \sum_{i=1}^{n} f_i \cdot t_i, \qquad b = \sum_{i=1}^{n} f_i \cdot (1 - t_i),$$

$$c = \sum_{i=1}^{n} (1 - f_i) \cdot t_i, \quad d = \sum_{i=1}^{n} (1 - f_i) \cdot (1 - t_i),$$

$$n = a + b + c + d$$

In our experiment, each "word" is an abstract noun, and n is the number of adjective types in the corpus (6407). Therefore, a indicates the number of adjective types co-occurring with both abstract nouns and b indicates the number of adjective types co-occurring only with the abstract noun corresponding to F. In contrast, c indicates the number of adjective types co-occurring only with the abstract noun corresponding to T and d indicates the number of adjective types that do not co-occur with either abstract noun.

3.2 Complementary similarity measure for gray-scale images

The CSM for gray-scale images is an extension of the CSM for binary images. Although the CSM for binary images is robust against graphical design, it is strongly affected by binarization or scanning conditions [Sawaki *et al.*, 1997].

The CSM for binary images is a special case of the four-fold point correlation coefficient. Therefore, Sawaki *et al.* [1997] defined the CSM for gray-scale images as a general form of the four-fold point correlation coefficient. Because it handles gray-scale images directly, this CSM is less affected by binarization or scanning conditions.

Let $F_g = (f_{g1}, f_{g2}, \ldots, f_{gi}, \ldots, f_{gn})$, where $f_{gi} = 0$ through 1, and let $T_g = (t_{g1}, t_{g2}, \ldots, t_{gi}, \ldots, t_{gn})$, where $t_{gi} = 0$ through 1, be the feature vectors of two gray-scale patterns. Then, the CSM_g of F_g to T_g is defined as follows:

$$CSM_g(F_g, T_g) = \frac{a_g d_g - b_g c_g}{\sqrt{n T_{g2} - T_g^2}}$$

$$a_g = \sum_{i=1}^{n} f_{gi} \cdot t_{gi}, \qquad b_g = \sum_{i=1}^{n} f_{gi} \cdot (1 - t_{gi}),$$

$$c_g = \sum_{i=1}^{n} (1 - f_{gi}) \cdot t_{gi}, \quad d_g = \sum_{i=1}^{n} (1 - f_{gi}) \cdot (1 - t_{gi}),$$

$$T_g = \sum_{i=1}^{n} t_{gi}, \quad T_{g2} = \sum_{i=1}^{n} t_{gi}^2$$

In our experiment, f_{gi} and t_{gi} are the weights based on the frequency at which an abstract noun co-occurred with an i-th type of adjective. In this paper, we used the following weighting function, where *Freq(noun, adj)* is the frequency at which the abstract noun co-occurs with the adjective:

$$Weight(noun, adj) = \frac{Freq(noun, adj)}{Freq(noun, adj) + 1} .$$

We paid particular attention to situations in which the noun co-occurred with the adjective. If the noun does not co-occur with the adjective, that is, *Freq*(*noun, adj*) is 0, the weight is 0.0. If *Freq*(*noun, adj*) is 1, it is 0.5. If *Freq*(*noun, adj*) was more than 1, the weight increased gradually until it approaches 1.0. This is because information on whether or not the noun co-occurs with the adjective is more important than information on how many times the noun co-occurs with the adjective.

4 Hierarchy Extraction Process

Word hierarchies were extracted as follows, where "TH" is a threshold value for each word pair under consideration;

1. Compute the similarity between appearance patterns for each pair of words. The hierarchical relation between the two words in a pair is determined by the similarity value. The pair is expressed as (X, Y), where X is a hypernym of Y and Y is a hyponym of X.
2. Sort the pairs by their normalized values and eliminate pairs with values below TH.
3. For each abstract noun C:
   - i Choose the hypernym-hyponym pair (C, D) with the highest value. This pair (C, D) is placed in the initial hierarchy.
   - ii Choose a pair (D, E) such that the hyponym E is not contained in the current hierarchy and (D, E) has the highest value among the pairs where word D, at the bottom of the current hierarchy, is a hypernym.

     Connect the hyponym E to D at the bottom of the current hierarchy.
   - iii Choose another pair (E, F) according to the previous step, and repeat the process until no more such pairs can be chosen.
   - iv Choose a pair (B, C) such that the hypernym B is not contained in the current hierarchy and (B, C) has the highest value among the pairs where the top word C of the current hierarchy is a hyponym.

     Connect the hypernym B in front of C at the top of the current hierarchy.
   - v Choose another pair (A, B) according to the previous step, and repeat the process until no more such pairs can be chosen.
4. If a short hierarchy is included in a longer hierarchy and the order of the words stays the same, the short one is dropped from the list of hierarchies.

5 Parameters

The conditions of our experiments were set as follows:

- CSM for binary images: TH = 0.2;
- CSM for gray-scale images: TH = 0.12.

If we set TH to a low value, it is possible to obtain long hierarchies. When the TH is too low, however, the number of word pairs that have to be considered becomes overwhelming and the reliability of the measurement decreases. We experimentally set TH as shown above so as to obtain "*koto* (matter)" as the top word of all hierarchies. Because "*koto*" co-occurred with the most number of adjectives, we predicted that "*koto*" would be at the top of the hierarchies.

6 Comparison and Evaluation

6.1 Overlap between extracted hierarchies

First, we compared the hierarchies obtained using the CSM for binary images (CSMb) and the CSM for gray-scale images (CSMg). Table 1 lists the number of extracted hierarchies. The CSMb extracted 189 hierarchies, while the CSMg extracted 178. There were only 28 common hierarchies with most of them having depths ranging from 3 to 6. One of the common hierarchies is shown below.

koto (matter) ---
  jyoutai (state) ---
    kankei (relation) ---
      tsunagari (relationship) ---
        en (ties/bonds).

The number of hierarchies obtained by the CSMg that included one or more hierarchies obtained by the CSMb was higher than the number obtained by the CSMb that included one or more obtained by the CSMg, i.e., (D) < (E). This suggests that the CSMg might be able to extract longer hierarchies than the CSMb can.

	Type of hierarchy	*Number*
(A)	Hierarchies obtained by CSMb	189
(B)	Hierarchies obtained by CSMg	178
(C)	Common hierarchies	28
(D)	Inclusion of CSMg hierarchies in CSMb hierarchies	5
(E)	Inclusion of CSMb hierarchies in CSMg hierarchies	38

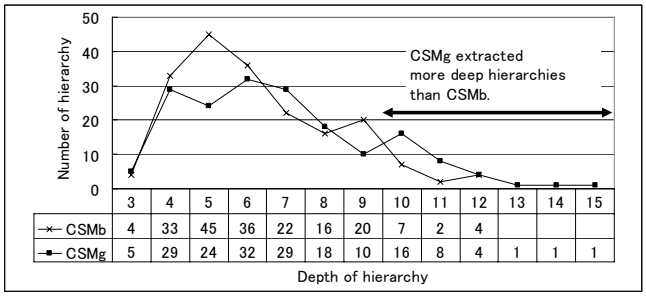
Table 1: Number of extracted hierarchies

The CSMg hierarchies shown below include one or more CSMb hierarchies. The underlined nouns are those that appear in one of the CSMb hierarchies.

koto (matter) ---
tokoro (point) ---
imeeji (image) ---
inshou (impression) ---
gaiken (appearance) ---
monogoshi (manner) ---
kihin (elegance) ---
hinkaku (grace) ---
kettou (pedigree) ---
kakei (family line)

koto (matter) ---
tokoro (point) ---
shigusa (behavior) ---
omokage (visage) ---
kawaisa (loveliness)

In fact, the depths of the CSMb hierarchies ranged from 3 to 12, while the depths of the CSMg hierarchies ranged from 3 to 15. We also found that the CSMg extracted more deep hierarchies than the CSMb but fewer shallow hierarchies. Overall, these results show that the CSMg extracted deeper hierarchies than did the CSMb, though the number of extracted hierarchies was smaller (see Figure 1).



Depth of hierarchy	3	4	5	6	7	8	9	10	11	12	13	14	15
CSMb	4	33	45	36	22	16	20	7	2	4			
CSMg	5	29	24	32	29	18	10	16	8	4	1	1	1

Figure 1: Distribution of hierarchies by depth

6.2 Agreement with the concept hierarchies in the EDR electronic dictionary

Next, we compared each of the hierarchies obtained by the CSMb and the CSMg with the concept hierarchies for adjectives in the EDR electronic dictionary.

In this paper, we extracted the hierarchies from corpora consisting mostly of newspaper articles. Because the newspaper articles cover a wide range of topics, the information we obtain from our corpora is general knowledge. Therefore, it is reasonable to compare our hierarchy of abstract nouns with existing general purpose hierarchies such as the EDR hierarchy.

The number of hierarchies for adjectives obtained from the EDR dictionary is 932, and the maximum depth is 14, whereas our hierarchies had maximum depths of 12 and 15, as noted above (Figure 1). Because both the EDR and our two types of hierarchies had similar maximum depths, it was appropriate to evaluate our hierarchies by comparing them with those of the EDR dictionary.

Hierarchy in the EDR electronic dictionary

The EDR electronic dictionary [1995], which was developed for advanced processing of natural language by computers, is composed of 11 sub-dictionaries, including a concept dictionary, word dictionaries, bilingual dictionaries, etc.

Although we could verify and analyse our extracted hierarchies by comparing them with the EDR dictionary, there were two problems with this approach. First, many concepts in the EDR dictionary are defined by sentences or phrases, whereas the concepts in our extracted hierarchies are defined by abstract nouns only. Therefore, we replaced the sentences and phrases in the EDR concept definitions with sequences of words for a more accurate comparison. Secondly, there was a difference of words between those used to define concepts in the EDR and the abstract nouns used for our extracted hierarchies. To solve this problem, we extracted synonyms from the EDR dictionary and added synonyms to both the words in the EDR concept definitions and the abstract nouns in our hierarchies. We thus transformed the conceptual hierarchies of adjectives in the EDR dictionary into hierarchies consisting of sequences of word sets to enable a comparison with our hierarchies consisting of adjective hypernyms.

Measurement of agreement level

For the comparison, we measured the degree of agreement with an EDR hierarchy for each extracted hierarchy, i.e., we counted the number of nodes that agreed with nodes in the corresponding EDR hierarchy, while maintaining the order of each hierarchy. A node in one of our hierarchies is a set of abstract nouns and their synonyms. We represented the node by *Node*(*abstract noun, synonym*$_1$*, synonym*$_2$*, ...*), where *Node* was a name identifying the node, while a node in an EDR hierarchy is a set of sequences of words and synonyms and was represented by *Node*(*content word1, synonym1*$_1$*, synonym1*$_2$*, ..., content word2, synonym2*$_1$*, synonym2*$_2$*, ...*). Therefore, if a word in a node of one of our hierarchies was included in a node of an EDR hierarchy, we considered that the nodes agreed. For example, if our hierarchy is

A(***a, a'***, *a"*) ---
B(***b***, *b'*, ***b"***) ---
C(*c, c', c"*) ---
D(***d***, *d', d"*)

and a corresponding EDR hierarchy is

P(***a, a'***, *x, x'*) ---
Q (***b, b"***) ---
R(*r, r', r"*) ---
S(*s, s'*, ***d***, *f, f', f"*) ---
T(*t, t', g, g'*),

we count three agreement nodes, because *A, B, D* match to *P, Q, S*, respectively. The bold words indicate words which match between our hierarchy and an EDR hierarchy. And we define the level of agreement as three.

In comparing hierarchies, we found cases in which a hypernym and its hyponym in our hierarchy were treated as synonyms in the EDR electronic dictionary. For example, consider the following hierarchy obtained using our approach:

koto (matter) ---
  tokoro (point) ---
    imeeji (image) ---
      funiki (atmosphere) ---
        kuuki (atmosphere in a place) ---
          kanjyou (feeling/emotion) ---
            shinjyou (one's feelings/one's sentiment) ---
              shinkyou (mental state/one's heart) ---
                kangai (deep emotion) ---
                  omoide (memories).

In the EDR electronic dictionary, each word is linked to its concept, and a synonymous relation is defined as words linked to the same concept. That is, we can gather synonyms via EDR dictionary. In fact, in the above hierarchy, we obtained "*shinjyou* (one's feelings/one's sentiment)" and "*shinkyou* (mental state/one's heart)" as synonyms of "*kanjyou* (feeling/emotion)" from the EDR dictionary. Also, we know that "*kuuki* (atmosphere in a place)" is a synonym of "*funiki* (atmosphere)".

If we count the agreement of the above hierarchy with the EDR dictionary strictly, the level of agreement is 6. The agreement nodes are "*koto* (matter) -- *tokoro* (point) -- *imeeji* (image) -- *funiki* (atmosphere) or *kuuki* (atmosphere in a place) -- *kanjyou* (feeling/emotion), *shinjyou* (one's feelings/one's sentiment), or *shinkyou* (mental state/one's heart) -- *omoide* (memories)". However, if we accept hypernym-hyponym relations among synonyms, the agreement level is 9. In this case, the agreement nodes are "*koto* (matter) -- *tokoro* (point) -- *imeeji* (image) -- *funiki* (atmosphere) -- *kuuki* (atmosphere in a place) -- *kanjyou* (feeling/emotion) -- *shinjyou* (one's feelings/one's sentiment) -- *shinkyou* (mental state/one's heart) -- *omoide* (memories)".

Results

Table 2 shows the distribution of CSMb hierarchies for various agreement levels. Table 3 shows the same results for the CSMg. The numbers in *italics* in the tables indicate that the number of hierarchies at that depth which are completely included in an EDR hierarchy. In the last column, we show the average agreement level at each depth. The value in the bottom right-hand corner is the average agreement level for all hierarchies.

Figure 2 is a graph of the average agreement level at each depth shown in Tables 2 and 3. In Figure 2, except at the depths of 8 and 9, the average agreement levels for the CSMg hierarchies are higher than those of the CSMb hierarchies. As shown in Tables 2 and 3, the deeper hierarchies tended to have higher agreement levels. Therefore, we consider that overall the CSMg hierarchies were closer to the EDR hierarchies than were the CSMb hierarchies. That is, the CSMg hierarchies were more in accordance with human intuition than were the CSMb hierarchies.

We also verified the ability of the CSM to estimate hypernym-hyponym relations between two nouns. Some of noun pairs whose relations were estimated by two CSMs were opposite each other. Table 4 shows some such pairs. For each of them, the CSMg estimated that the noun on the left was a hypernym and the one on the right was a hyponym. The CSMb estimated the reverse.

Depth of hierarchy	*Agreement level* 1	2	3	4	5	6	7	8	9	*Ave. level*
3	1	2	***1***							2.00
4		6	18	***9***						3.09
5		7	23	12	***3***					3.24
6		4	12	9	7	***4***				3.86
7		2	2	10	4	3	***1***			4.32
8			1	6	6	3				4.69
9			1	4	5	4	5	1		5.55
10			2		2	2			1	5.29
11			1			1				4.50
12			1	1					2	6.25
Overall ave.										4.28

Table 2: Distribution of CSMb hierarchies for various agreement levels

Depth of hierarchy	*Agreement level* 1	2	3	4	5	6	7	8	9	*Ave. level*
3	1	3	***1***							2.50
4		6	13	***10***						3.14
5		3	9	9	***3***					3.50
6		1	11	12	6	***2***				3.91
7		1	5	10	8	5				4.38
8			4	5	7	2				4.39
9				6	1	3				4.70
10					2	6	4	3	1	6.69
11			1	2	1	3	1			5.13
12								1	3	8.75
13							1			7.00
14								1		8.00
15									1	9.00
Overall ave.										5.47

Table 3: Distribution of CSMg hierarchies for various agreement levels

Noun pair	
tokoro (point)	***imeeji*** (image)
tokoro (point)	***men*** (side)
tokoro (point)	***inshou*** (impression)
tokoro (point)	***seikaku*** (character)
tokoro (point)	***seishitsu*** (property)
tokoro (point)	***kanshoku*** (touch)
kimochi (thought/feeling/intention)	***omoi*** (feeling/mind/expectation)
kagayaki (brightness)	***koutaku*** (gloss)
kuukan (space)	***men*** (side)
kotoba (speech)	***iken*** (opinion)
kokoro (mind)	***shinjyou*** (one's feelings/one's sentiment)
hiyori (fine weather)	***ondo*** (temperature)

Table 4: Noun pairs estimated oppositely

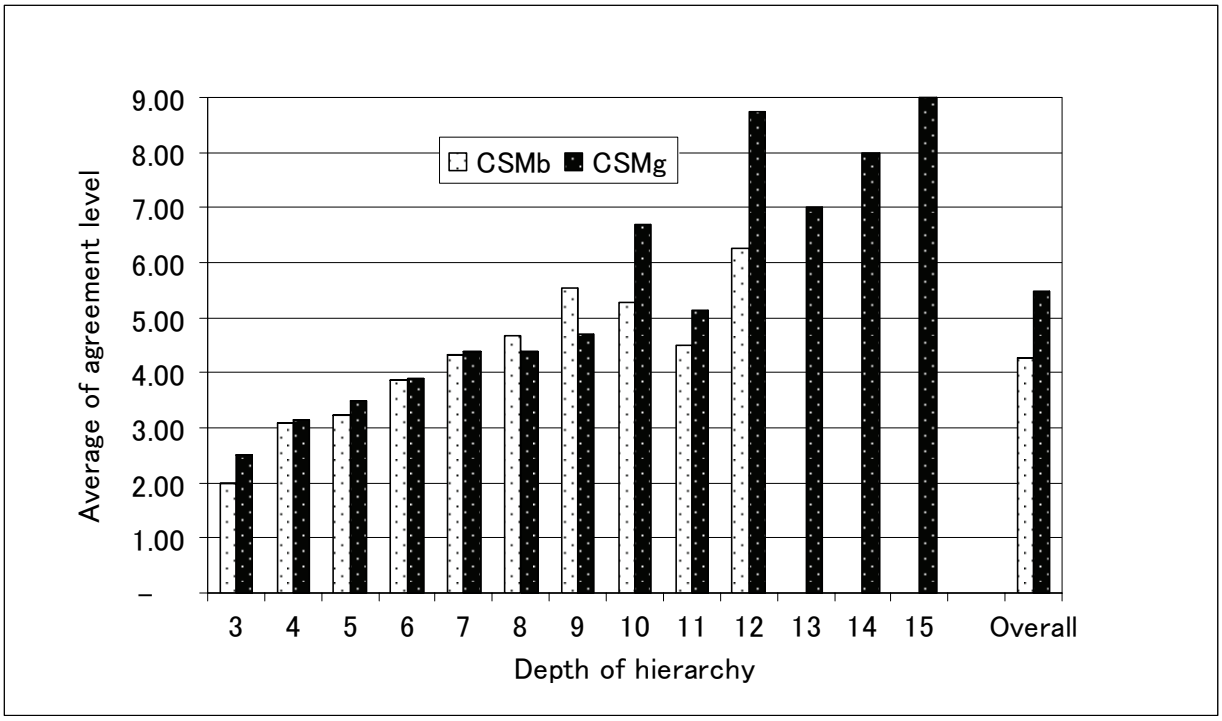


Figure 2: Comparison of CSMb and CSMg hierarchies by average agreement level

In our experiment, there were 836 such pairs. As the total number of pairs considered was 17201, these pairs amounted to less than 5%. We also found that in most cases these pairs appeared in the middle of a hierarchy.

Let us consider the hypernym-hyponym relation between "*kimochi* (thought/feeling/intention)" and "*omoi* (feeling/mind/expectation)". A CSMg hierarchy including "*kimochi*" and "*omoi*" was as follows:

koto (matter) ---
tokoro (point) ---
imeeji (image) ---
inshou (impression) ---
kanji (feeling/sense) ---
kibun (feeling/mood) ---
kimochi (thought/feeling/intention) ---
omoi (feeling/mind/expectation) ---
negai (wish) ---
nen (desire).

Here, "*kimochi*" was estimated as a hypernym of "*omoi*." However, CSMb estimated the opposite, i.e., "*omoi*" was a hypernym of "*kimochi*." We examined the values given by the CSMb and CSMg to "*kimochi*" and "*omoi*" (see Table 5).

(*F, T*)	("*omoi*", "*kimochi*")	("*kimochi*", "*omoi*")	*Diff.*
CSMb	0.8094	0.8064	0.0030
CSMg	0.7632	0.7700	0.0068

Table 5: Differences in CSM values for "*omoi*" and "*kimochi"*

As shown in Table 5, the values of CSMg in both directions were similar, as was the case for CSMb. As the CSM is a measure of inclusion, if the values of the CSM for two words in both directions are similar, it might mean that the two words are synonymous. Both the CSMg and CSMb estimated that "*kimochi*" and "*omoi*" were similar. However, due to the very small differences in values, they extracted opposite results. In fact, in the EDR electronic dictionary, "*kimochi*" and "*omoi*" are synonymous because both have the meaning of "feeling". The pairs estimated oppositely by the two CSMs may have a synonymous relation. In future work, we will introduce the CSM-based definition to estimate two words as synonyms.

A small number of word pairs were estimated oppositely by the two CSMs, with large differences in the CSM value of each direction. For example, CSMg estimated that "*tokoro* (point)" was a hypernym of "*imeeji* (image)" and CSMb estimated that "*imeeji*" was a hypernym of "*tokoro*" (see Table 6).

(*F, T*)	("*tokoro*", "*imeeji*")	("*imeeji*", "*tokoro*")	*Diff.*
CSMb	0.6767	0.7156	+0.0389
CSMg	0.6631	0.6468	-0.0163

Table 6: Differences in CSM values for "*tokoro*" and "*imeeji"*

The introduction of frequency information for CSMg may be the reason for this difference. In future work, we will analyze this in more detail.

7 Conclusion

We proposed a method of automatically extracting hierarchies based on the inclusion relations of the appearance patterns of words from corpora. In this paper, we described our attempts to extract objective hierarchies of abstract nouns co-occurring with adjectives in Japanese. We applied the complementary similarity measure for gray-scale images (CSMg) to search for better hierarchical word structures. In our experiment, we found that the CSMg could extract word hierarchies from corpora, even though it was developed for recognizing degraded machine-printed text. We also compared the CSMg with the CSM for binary images (CSMb), and found that the CSMg hierarchies were more in accordance with human intuition than were the CSMb hierarchies, as measured by their degree of agreement with the EDR electronic dictionary. As a next step, it would be interesting to compare this method to existing statistical methods such as agglomerative clustering. It would also be necessary to estimate confidence.

In this paper, we thus verified the suitability of the proposed method for extracting hierarchies from corpora. We consider that hierarchies tuned to specific corpora could be used for query expansion in information retrieval for specific domains. Our future work will include extracting hierarchies of key words in corpora and trying to utilize them as a special thesaurus for domain-oriented information retrieval.

References

[Berland and Charniak, 1999] Berland, M. and Charniak, E. Finding parts in very large corpora, In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 57-64, 1999.

[Caraballo, 1999] Caraballo, S. A. Automatic construction of a hypernym-labeled noun hierarchy from text, In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 120-126, 1999.

[EDR, 1995] *EDR Electronic Dictionary.* 1995.

http://www2.nict.go.jp/kk/e416/EDR/index.html

[Hagita and Sawaki, 1995] Hagita, N. and Sawaki, M. Robust recognition of degraded machine-printed characters using complementary similarity measure and error-correction learning, In *Proceedings of the SPIE – The International Society for Optical Engineering*, 2442: pp. 236-244, 1995.

[Hearst, 1992] Hearst, M. A. Automatic acquisition of hyponyms from large text corpora, In *Proceedings of the 14th International Conference on Computational Linguistics*, pp. 539-545, 1992.

[Kanzaki *et al.*, 2003] Kanzaki, K., Ma, Q., Yamamoto, E., Murata, M. and Isahara, H. Adjectives and their abstract concepts --- toward an objective thesaurus from semantic map. In *Proceedings of the Second International Workshop on Generative Approaches to the Lexicon*, pp. 177-184, 2003.

[Kanzaki *et al*., 2004] Kanzaki, K., Yamamoto, E., Ma, Q. and Isahara, H. Construction of an objective hierarchy of abstract concepts via directional similarity. In *Proceedings of 20th International Conference on Computational Linguistics,* Vol. II , pp. 1147-1153, 2004.

[Matsumoto *et al.*, 1996] Matsumoto, Y., Sudo, S., Nakayama, T. and Hirao, T. Thesaurus construction from multiple language resources, In *IPSJ SIG Notes NL-93*, pp. 23-28, 1996.

[Miller *et al.*, 1990] Miller, A., Beckwith, R., Fellbaum, C., Gros, D., Millier, K. and Tengi, R. *Five Papers on WordNet, Technical Report CSL Report 43*, Cognitive Science Laboratory, Princeton University, 1990.

[Nakayama and Matsumoto, 1997] Nakayama, T. and Matsumoto, Y. Positioning nouns in a classification-based thesaurus, In *IPSJ SIG Notes NL-120*, pp. 103-108, 1997.

[Nemoto, 1969] Nemoto, K. Combination of noun with "ga-case" and adjective, *Language Research for the Computer 2*, National Language Research Institute, pp. 63-73, 1969 (In Japanese).

[Sawaki *et al.*, 1997] Sawaki, M., Hagita, N. and Ishii, K. Robust character recognition of gray-scaled images with graphical designs and noise, In *Proceedings of the International Conference on Document Analysis and Recognition,* pp. 491-494, 1997.

[Shoutsu *et al.*, 2003] Shoutsu, Y., Tokunaga, T. and Tanaka, H. The integration of Japanese dictionary and thesaurus, In *IPSJ SIG Notes NL-153*, pp. 141-146, 2003.

[Takahashi, 1975] Takahashi, T. Various phase related to part-whole relation investigated in sentence, *Studies in the Japanese Language 103*, The Society of Japanese Linguistics, pp. 1-16, 1975 (In Japanese).

[Tsurumaru *et al.*, 1986] Tsurumaru, H., Hitaka, T. and Yoshida, S. Automatic extraction of hierarchical relation between words, In *IPSJ SIG Notes NL-83*, pp. 121-128, 1986.

[Yamamoto and Umemura, 2002] Yamamoto, E. and Umemura, K. A similarity measure for estimation of one–to-many relationship in corpus, In *Journal of Natural Language Processing*, pp. 45-75, 2002.

[Yamamoto *et al.*, 2004] Yamamoto, E., Kanzaki, K. and Isahara, H. Hierarchy extraction based on inclusion of appearance, In *ACL04 Companion Volume to the Proceedings of the Conference*, pp. 149-152, 2004.