Hierarchy Extraction based on Inclusion of Appearance

Eiko Yamamoto **Kyoko Kanzaki** **Hitoshi Isahara**

Computational Linguistics Group,
National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Soraku-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 relation of appearance patterns from corpora. We apply a complementary similarity measure to find a hierarchical word structure. This similarity measure was developed for the recognition of degraded machine-printed text in the field and can be applied to estimate one-to-many relations. Our purpose is to extract word hierarchies from corpora automatically. As the initial task, we attempt to extract hierarchies of abstract nouns co-occurring with adjectives in Japanese and compare with hierarchies in the EDR electronic dictionary.

1 Introduction

The hierarchical relations of words are useful as language resources. Hierarchical semantic lexical databases such as WordNet (Miller et al., 1990) and the EDR electronic dictionary (1995) are used for NLP research worldwide to fully understand a word meaning. In current thesauri in the form of hierarchical relations, words are categorized manually and classified in a top-down manner based on human intuition. This is a good way to make a lexical database for users having a specific purpose. However, word hierarchies based on human intuition tend to vary greatly depending on the lexicographer. In addition, hierarchical relations based on various data may be needed depending on each user.

Accordingly, we try to extract a hierarchical relation of words automatically and statistically. In previous research, ways of extracting from definition sentences in dictionaries (Tsurumaru et al., 1986; Shoutsu et al., 2003) or from a corpus by using patterns such as “a part of”, “is-a”, or “and” (Berland and Charniak, 1999; Caraballo, 1999) have been proposed. Also, there is a method that uses the dependence relation between words taken from a corpus (Matsumoto et al., 1996). In contrast, we propose a method based on the inclusion relation of appearance patterns from corpora.

In this paper, to verify the suitability of our method, we attempt to extract hierarchies of abstract nouns co-occurring with adjectives in Japanese. We select two similarity measures to estimate the inclusion relation between word appearance patterns. One is a complementary similarity measure; i.e., a similarity measure developed for the recognition of degraded machine-printed text in the field (Hagita and Sawaki, 1995). This measure can be used to estimate one-to-many relations such as superordinate–subordinate relations from appearance patterns (Yamamoto and Umemura, 2002). The second similarity measure is the overlap coefficient, which is a similarity measure to calculate the rate of overlap between two binary vectors. Using each measure, we extract hierarchies from a corpus. After that, we compare these with the EDR electronic dictionary.

2 Experiment Corpus

A good deal of linguistic research has focused on the syntactic and semantic functions of abstract nouns (Nemoto, 1969; Takahashi, 1975; Schmid, 2000; Kanzaki et al., 2003). In the example, “*Yagi* (goat) *wa seishitsu* (nature) *ga otonashii* (gentle) (The nature of goats is gentle).”, Takahashi (1975) recognized that the abstract noun “*seishitsu* (nature)” is a hypernym of the attribute that the predicative adjective “*otonashi* (gentle)” expresses. Kanzaki et al. (2003) defined such abstract nouns that co-occur with adjectives as adjective hypernyms, and extracted these co-occurrence relations between abstract nouns and adjectives from many corpora such as newspaper articles. In the linguistic data, there are sets of co-occurring adjectives for each abstract noun – the total number of abstract noun types is 365 and the number of adjective types is 10,525. Some examples are as follows.

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

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

3 Complementary Similarity Measure

The complementary similarity measure (CSM) is used in a character recognition method for binary images which is robust against heavy noise or graphical designs (Sawaki and Hagita, 1996). Yamamoto et al. (2002) applied CSM to estimate one-to-many relations between words. They estimated one-to-many relations from the inclusion relations between the appearance patterns of two words. The appearance pattern is expressed as an n-dimensional binary feature vector. Now, let F = (f_1, f_2, …, f_n) and T = (t_1, t_2, …, t_n) (where f_i, t_i = 0 or 1) be the feature vectors of the appearance patterns for a word and another word, respectively. The CSM of F to T is defined as

$$CSM(F,T) = \frac{ad - bc}{\sqrt{(a+c)(b+d)}}$$
$$a = \sum_{i=1}^{n} f_i \cdot t_i, \qquad b = \sum_{i=1}^{n} (1 - f_i) \cdot t_i,$$
$$c = \sum_{i=1}^{n} f_i \cdot (1 - t_i), \quad d = \sum_{i=1}^{n} (1 - f_i) \cdot (1 - t_i),$$
$$n = a + b + c + d$$

The CSM of F to T represents the degree to which F includes T; that is, the inclusion relation between the appearance patterns of two words.

In our experiment, each "word" is an abstract noun. Therefore, *n* is the number of adjectives in the corpus, *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, and *d* indicates the number of adjectives co-occurring with neither abstract noun.

4 Overlap Coefficient

The overlap coefficient (OVLP) is a similarity measure for binary vectors (Manning and Schutze, 1999). OVLP is essentially a measure of inclusion. It has a value of 1.0 if every dimension with a non-zero value for the first vector is also non-zero for the second vector or vice versa. In other words, the value is 1.0 when the first vector completely includes the second vector or vice versa. OVLP of F and T is defined as

$$OVLP(F,T) = \frac{|F \cap T|}{MIN(|F|, |T|)} = \frac{a}{MIN(a+b, a+c)}$$

5 EDR hierarchy

The EDR Electronic Dictionary (1995) was developed for advanced processing of natural language by computers and is composed of eleven sub-dictionaries. The sub-dictionaries include a concept dictionary, word dictionaries, bilingual dictionaries, etc. We verify and analyse the hierarchies that are extracted based on a comparison with the EDR dictionary. However, the hierarchies in EDR consist of hypernymic concepts represented by sentences. On the other hand, our extracted hierarchies consist of hypernyms such as abstract nouns. Therefore, we have to replace the concept composed of a sentence with the sequence of the words. We replace the description of concepts with entry words from the "Word List by Semantic Principles" (1964) and add synonyms. We also add to abstract nouns in order to reduce any difference in representation. In this way, conceptual hierarchies of adjectives in the EDR dictionary are defined by the sequence of words.

6 Hierarchy Extraction Process

The processes for hierarchy extraction from the corpus are as follows. "TH" is a threshold value for each pair under consideration. If TH is low, we can obtain long hierarchies. However, if TH is too low, the number of word pairs taken into consideration increases overwhelmingly and the measurement reliability diminishes. In this experiment, we set 0.2 as TH.

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. We express the pair as (X, Y), where X is a hypernym of Y and Y is a hyponym of X.
2. Sort the pairs by the normalized similarities and reduce the pairs where the similarity is less than TH.
3. For each abstract noun,
   A) Choose a pair (B, C) where word B is the hypernym with the highest value. The hierarchy between B and C is set to the initial hierarchy.
   B) Choose a pair (C, D) where hyponym D is not contained in the current hierarchy and has the highest value in pairs where the last word of the current hierarchy C is a hypernym.
   C) Connect hyponym D with the tail of the current hierarchy.
   D) While such a pair can be chosen, repeat B) and C).
   E) Choose a pair (A, B) where hypernym A is not contained in the current hierarchy and has the highest value in pairs where the first word of the current hierarchy B is a hypernym.
   F) Connect hypernym A with the head of the current hierarchy.
   G) While such a pair can be chosen, repeat E) and F).

4. For the hierarchies that are built,
   A) If a short hierarchy is included in a longer hierarchy with the order of the words preserved, the short one is dropped from the list of hierarchies.
   B) If a hierarchy has only one or a few different words from another hierarchy, the two hierarchies are merged.

7 Extracted Hierarchy

Some extracted hierarchies are as follows. In our experiment, we get ***koto*** (matter) as the common hypernym.

koto (matter) -- ***joutai*** (state) -- ***kankei*** (relation) -- ***kakawari*** (something to do with) -- ***tsukiai*** (have an acquaintance with)

koto (matter) -- ***toki*** (when) -- ***yousu*** (aspect) -- ***omomochi*** (one's face) -- ***manazashi*** (a look) -- ***iro*** (on one's face) -- ***shisen*** (one's eye)

8 Comparison

We analyse extracted hierarchies by using the number of nodes that agree with the EDR hierarchy. Specifically, we count the number of nodes (nouns) which agree with a word in the EDR hierarchy, preserving the order of each hierarchy. Here, two hierarchies are "A - B - C - D - E" and "A - B - D - F - G." They have three agreement nodes; "A - B - D."

Table 1 shows the distribution of the depths of a CSM hierarchy, and the number of nodes that agree with the EDR hierarchy at each depth. Table 2 shows the same for an OVLP one. "Agreement Level" is the number of agreement nodes. The bold font represents the number of hierarchies completely included in the EDR hierarchy.

8.1 Depth of Hierarchy

The number of hierarchies made from the EDR dictionary (EDR hierarchy) is 932 and the deepest level is 14. The number of CSM hierarchies is 105 and the depth is from 3 to 14 (Table 1). The number of OVLP hierarchies is 179 and the depth is from 2 to 9 (Table 2). These results show that CSM builds a deeper hierarchy than OVLP, though the number of hierarchies is less than OVLP. Also, the deepest level of CSM equals that of EDR. Therefore, comparison with the EDR dictionary is an appropriate way to verify the hierarchies that we have extracted.

In both tables, we find most hierarchies have an agreement level from 2 to 4. The deepest agreement level is 6. For an agreement level of 5 or better, the OVLP hierarchy includes only two hierarchies while the CSM hierarchy includes nine hierarchies. This means CSM can extract hierarchies having more nodes which agree with the EDR hierarchy than is possible with OVLP.

Depth of Hierarchy	*Agreement Level*					
	1	*2*	*3*	*4*	*5*	*6*
3	1	4	**1**			
4		8	6	**2**		
5		9	8		**1**	
6		8	9	4	1	
7		2	6	1	1	
8		1	5	2	2	
9		3	2	3		1
10			1		2	
11			4	1		
12			1			1
13			1			2
14					1	

Table 1: Distribution of CSM hierarchy for each depth

Depth of Hierarchy	*Agreement Level*					
	1	*2*	*3*	*4*	*5*	*6*
2		**1**				
3	2	8	**1**			
4		25	9	**1**		
5		24	13	7		
6		21	31	5		
7		5	12	1		1
8		3	5	2	1	
9		1	3	1		

Table 2: Distribution of OVLP hierarchy for each depth

Also, many abstract nouns agree with the hyperonymic concept around the top level. In current thesauri, the categorization of words is classified in a top-down manner based on human intuition. Therefore, we believe the hierarchy that we have built is consistent with human intuition, at least around the top level of hyperonymic concepts.

9 Conclusion

We have proposed a method of automatically extracting hierarchies based on an inclusion relation of appearance patterns from corpora. In this paper, we attempted to extract objective hierarchies of abstract nouns co-occurring with adjectives in Japanese. In our experiment, we showed that complementary similarity measure can extract a kind of hierarchy from corpora, though it is a similarity measure developed for the recognition of degraded machine-printed text. Also, we can find interesting hierarchies which suit human intuition, though they are different from exact hierarchies. Kanzaki et al. (2004) have applied our approach to verify

classification of abstract nouns by using self-organization map. We can look a suitability of our result at that work.

In our future work, we will use our approach for other parts of speech and other types of word. Moreover, we will compare with current alternative approaches such as those based on sentence patterns.

References

Berland, M. and Charniak, E. 1999. Finding Parts in Very Large Corpora, In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp.57-64.

Caraballo, S. A. 1999. 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.

EDR Electronic Dictionary. 1995.

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

Hagita, N. and Sawaki, M. 1995. 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.

Kanzaki, K., Ma, Q., Yamamoto, E., Murata, M., and Isahara, H. 2003. 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.

Kanzaki, K., Ma, Q., Yamamoto, E., Murata, M., and Isahara, H. 2004. Extraction of Hyperonymy of Adjectives from Large Corpora by using the Neural Network Model. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Volume II, pp.423-426.

Kay, M. 1986. *Parsing in Functional Unification Grammar*. In "Readings in Natural Language Processing", Grosz, B. J., Spark Jones, K. and Webber, B. L., ed., pp.125-138, Morgan Kaufmann Publishers, Los Altos, California.

Manning, C. D. and Schutze, H. 1999. Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge MA.

Matsumoto, Y. and Sudo, S., Nakayama, T., and Hirao, T. 1996. Thesaurus Construction from Multiple Language Resources, In *IPSJ SIG Notes NL-93*, pp.23-28 (In Japanese).

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

Mosteller, F. and Wallace, D. 1964. *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, Reading, Massachusetts.

Nemoto, K. 1969. The combination of the noun with "ga-Case" and the adjective, Language research2 for the computer, National Language Research Institute, pp.63-73 (In Japanese).

Shmid, H-J. 2000. *English Abstract Nouns as Conceptual Shells*, Mouton de Gruyter.

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

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1): pp.11-21.

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, pp.1-16 (In Japanese).

Tsurumaru, H., Hitaka, T., and Yoshita, S. 1986. Automatic extraction of hierarchical relation between words, In *IPSJ SIG Notes NL-83*, pp.121-128 (In Japanese).

Yamamoto, E. and Umemura, K. 2002. A Similarity Measure for Estimation of One–to-Many Relationship in Corpus, In *Journal of Natural Language Processing*, pp.45-75 (In Japanese).

Word List by Semantic Principles. 1964. National Language Research Institute Publications, Shuei Shuppan (In Japanese).