Automatic Extraction of Hyponyms from Japanese Newspapers Using Lexico-syntactic Patterns

Maya Ando[1)], Satoshi Sekine[2)] and Shun Ishizaki[1)]

1) Graduate School of Media and Governance, Keio University
5322 Endo, Fujisawa-shi, Kanagawa, 252-8520, Japan
{maya, ishizaki}@sfc.keio.ac.jp

2) Computer Science Department, New York University
715 Broadway, 7th floor, New York, NY 10003 USA
sekine@cs.nyu.edu

Abstract

We describe a method to automatically extract hyponyms from Japanese newspapers. First, we discover patterns which can extract hyponyms of a noun, such as "A nado-no B (B such as A)", then we apply the patterns to the newspaper corpus to extract instances. The procedure works best to extract hyponyms of concrete things in the middle of the word hierarchies. The precision is 49-87 percent depending on the patterns. We compare the extracted hyponyms and those associated by humans. We find that the popular words in the associative concept dictionary are likely to be found in the corpus but also many additional hyponyms can be extracted from 32 years of newspaper articles.

Introduction

Semantic relationships between words are important in advanced natural language processing. There are several large databases describing the relationship between words in Japanese. The *Goitaikei* developed by NTT contains 300,000 Japanese words (*Nihongo Goitaikei* 2002) and EDR electric dictionary (EDR homepage) contains 410,000 concepts. These are general dictionaries, but do not include frequency or other significant information, i.e. which words are more strongly related to its parent and so on. Recently an associative concept dictionary (Okamoto 2001) was created. It is a list of words associated by human subjects for about 1000 stimulus words based on seven relationships (hypernym, hyponym, part/material, attribute, synonym, action/ related and situation/included). The information includes the association time, the order of associated words, and how many people made this association. So, this associative concept dictionary includes significant information, however it is not so easy to scale up.

In this paper we will describe a method to automatically extract hyponyms of nouns from Japanese newspapers using lexico-syntactic patterns from a large corpus. One of the motivations is to see if the corpus method can be a supplement to the hyponym lists associated by human. A very closely related previous study was conducted by Hearst (1992; 1998) in Englsih. She created lexico-syntactic patterns, e.g. "A such as B" and applied them to a corpus to extract hyponym words. She reported that the precision was 63%. We apply a similar method in Japanese and extract hyponyms from a Japanese corpus. Then we compare the results with the associative concept dictionary, i.e. the list of words associated by human.

The overall procedure can be divided into three tasks; 1) to select lexico-syntactic patterns for the hyponym relationship, 2) to extract hyponyms from a Japanese corpus using the patterns, and 3) to evaluate it by comparing the output with the associative concept dictionary.

Discovery of Lexicosyntactic Patterns

We first investigate patterns that include some known pairs of a word and its hyponym in 6 years newspaper corpus. The patterns are expressed in terms of Japanese dependency relationships within a certain distance. We discovered about 30 candidate patterns and selected 7 patterns which have significant frequencies and accuracies, shown in Table 1. The patterns we did not select include "A *wo hajime* B", (A as an example of B) or "A *wo fukume* B" (B including A).

	Japanese patterns	English Translation
p1	A *nado* B	B such as A
p2	A *nado-no* B	
p3	A *ni-nita* B	B which is similar to A
p4	A *no-youna* B	B which look/sound like A
p5	A *to-iu* B	B which is called/named A
p6	A *to-yoba-reru* B	
p7	A *igai-no* B	B other than A

Table1: Lexico-syntactic Patterns

patterns	frequency	precision		
		Concrete	Abstract	Total
p1	10277	42%	15%	33%
p2	16463	53%	22%	43%
p3	448	76%	38%	67%
p4	3160	46%	17%	36%
p5	14719	50%	23%	40%
p6	481	73%	46%	65%
p7	1151	84%	25%	70%

Table 2: Frequency and precision for patterns

Extraction of Hyponyms from Newspaper

We applied these seven patterns to 32 years of newspaper (different from the 6 years of newspaper used in the pattern discovery) to find hyponyms of 130 hypernyms used in the associative concept dictionary. For example, in order to extract hyponyms of "*yasai* (vegetable)" we apply the pattern "XXX *nado-no yasai* (vegetables such as XXX)" to the corpus. Then we extract "tomato" as a hyponym when we discover the phrase "*tomato nado-no yasai* (vegetables such as a tomato)". Table 2 shows the frequency of 7 patterns in the corpus and the precision of the extracted hyponyms for each pattern depending on the hypernym types, concrete things, abstract things and total.

The precision was calculated based on up to 30 sample hyponyms for each hypernym. We regard A as a correct hyponym of B, if the pair of words make sense in the phrase "A *ha* B *dearu* (A is a B)" without any context. Table 3 shows the result of the evaluation for some hypernyms. It shows frequencies of each hypernym with particular pattern (frequency) and number of different hyponyms found with the hypernym and the pattern (type). Note that these frequencies include the incorrect hyponyms. The precision was calculated on up to 30 samples for each data. In the rest of this section, we will discuss four interesting observations about these results.

Type of hypernym

In general, concrete things have higher precision than abstract things, as shown in Table 2 and Table 3. For example, the average precision of "*norimono* (general sense of vehicle including vessel and airplane)", which has hyponyms such as "car", "train" and "airplane", is 80% even with high frequency of 137. Although the average precision of "*basu"*(bus) is 18%, the precision of abstract things is quite low. For example the average precision of "*imi* (meaning)" is 0% and "*ryokou* (travel)" is 3%. In the associative concept dictionary some abstract things have a few hyponyms such as "*kaigai-ryokou* (foreign travel)", and "*shinkon-ryokou* (honeymoon)" for "*ryokou* (travel)"; however, in the corpus those hyponyms do not appear much. Also some patterns are used with different meanings, especially for abstract things. For example, p2 is usually used to itemize examples but it sometimes expresses possession, e.g. "*kaisha nado-no ryokou* (travel held by a company)". Also, p5 "*A to-iu B*"is sometime used in the context like "*C to-ha A to-iu B* (B of C is A)". For example, "*kaosu to-ha konton to-iu imi* (the meaning of "chaos" is *konton (*a Japanese word for chaos))". That is to say, *konton* is content of meaning. These kinds of usages are observed more often if the hypernym is an abstract thing.

Hypernym's location in concept hierarchy

We found that the precision depends on whether the hypernym is a relatively general term (located at middle position of general concept hierarchy) or a specific term (located at lower position of hierarchy). The precision for words in the middle of concept hierarchies is relatively high. On the other hand, for words lower in the concept hierarchy, it can hardly find any hyponyms and even if it finds some, the precision is quite low. For example the precision of words in the middle of the concept hierarchy such as "*norimono* (vehicle)", "*doubutsu* (animal)" and "*kagu* (furniture)" is more than 80%. On the other hand, for words lower in the concept hierarchy such as "*basu*(bus)", "*inu*(dog)" and "*isu*(chair)", the precision is less than 50%.

Different patterns

We found that the precision of extracted hyponyms and their characteristics depend on the patterns.

When a hypernym is a concrete thing, pattern p4 tends to have a lower precision than other patterns, because P4 can be a figurative expression. For example, in expression "*omocha no-youna gakki* (a music instrument which looks like a toy)", "music instrument" is not a hyponym of "toy". However, the precision of p4 is higher than p1 and p2, when a hypernym is a kind of a human such as "*ningen* (human)", "*sensei* (teacher)" and "*josei* (woman)". P4 is more likely to occur with a proper noun than p1 and p2. Because "*ningen* (human)" has more proper nouns as hyponyms than "*norimono* (vehicle)" and "*doubutsu* (animal)", it appears frequently in p4 and the accuracy is high.

P6 can be used in order to introduce not so popular nouns for the category. For example, we can usually find a phrase such as "*arupaka toiu doubutsu* (an animal which is called an alpaca)" Or "*Naomi toiu josei* (a woman whose name is Naomi)". In both cases, hyponyms are found which are relatively rare words or proper nouns.

Dependency analysis error

We found that there are typical errors caused by dependency analysis errors. For example, in the sentence "*kanzo nado-no ningen no zouki*" (internal organs of human such as liver), "liver" has to depend on "organs", rather than "human". This is an interesting problem, as the ambiguity may have to be solved by the knowledge we are trying to achieve, i.e. the hypernym and hyponym relationship among "liver", "organ" and "human". This is a chicken-and-egg problem, but as we can use a large corpus, we hope that some sophisticated statistical methods can solve the problem (Sekine et. al 1992).

Comparison with Human Association

It would be interesting to see how different (or similar) the hyponyms created by human are from those extracted from a corpus using the proposed method. In this section we compare the hyponyms extracted from a corpus (extracted hyponyms) against the hyponyms associated by humans (associated hyponyms). In order to minimize the noise, we restrict the associated hyponyms to those associated by more than 2 subjects, as singletons include many peculiar and sometime erroneous words. Note that the number of subjects for the association experiment was 50.

Three kinds of investigations were conducted mainly on those hyponyms which have high frequencies and high precision. As a result, the investigated hypernyms are mostly concrete things in the middle of the word hierarchies explained in the previous section.

Overlap between extracted & associated hyponyms

For the examined data, about 55% (442/810) of associated hyponyms are found among the extracted hyponyms, and 18% (143/786) of the correct extracted hyponym are found in the associated hyponyms. Table 4 shows the overlap between extracted hyponyms and associated hyponyms for 7 hypernyms. The coverage for each hyponyms is also shown. The overlap hyponyms are the 3 most frequently extracted hyponyms found in both the associated and extracted hyponyms, in the order of popularity of association. "Not extracted hyponyms" are hyponyms in associated dictionary, but not extracted in the experiment, also in the order of popularity of association. The 3 most popular associated hyponyms are shown. "Not associated hyponyms" are hyponyms not in associated dictionary, but extracted in the experiment. The 3 most frequent extracted hyponyms are shown.

In general, hypernyms located lower in the general concept hierarchy have few hyponyms in both association and extraction. For example "tora(tiger)" has 3 associated hyponyms such as "white tiger" and there are no extracted hyponyms.
Sometimes a difference of focus the difference between extracted and associated hyponyms. "*norimono* (vehicle)" usually occur as a traffic facility or industry in newspaper and "Train", "car", and "airplane" are extracted. But subjects associate words that they can see in daily life such as "tricycle" or "horse".
For some hypernyms, there are hyponyms, which include the literal string of the hypernym. For example, "basu (bus)" has hyponyms like "*torori-basu* (trolleybus)", "*kankou basu* (sightseeing bus)" or "*junkai basu* (loop-line bus)". We found that these hyponyms tend not to be extracted, even though people can easily associate. It is because expressions like "A bus such as a loop-line bus" are a bit awkward and do not appear in the corpus a lot.
For the hypernym "*ningen* (human)", the types of associated hyponyms and extracted hyponyms are quite different. Almost all the associated words are general nouns such as "Japanese" and "teacher", but many extracted words are pronouns and proper nouns, such as "I", "you" and "Beethoven".
In general, extracted hyponyms are more frequently nouns at low positions in hierarchies, such as "dog" and "cat" as hyponyms of "*doubutsu* (animal)", rather than nouns in the middle of hierarchies such as "amphibian" and "flesh-eating animal". We believe this is because the extraction patterns are likely to be used with a concrete example.

How much more we found over the association

In table 4, we showed examples of hyponyms extracted from corpora, but not associated by subjects. This is one of the purposes of our experiments, that is to see if a corpus-based method can be a supplement to human created knowledge. So we count up how many additional hyponyms are extracted from the corpus over the associative concept dictionary. The system finds 10% to 320% additional hyponyms over associated hyponyms (average 89%). For example, for "*doubutsu* (animal)", we found 120 additional hyponyms such as "*buta* (pig)", "*kaeru* (frog)" and "*azarashi*(seal)", which was not in the associated list for animals of 38 instances.

Popularity in associated dictionary

We examined the relationship between the popularity in the associative concept dictionary (the number of subjects who associated the hyponym) and whether the hyponym is extracted from the corpus. The associated hyponyms are divided into 3 groups; the words associated by more than 20% of subjects as Group 1, 10-20% as Group2 and 2-10% as Group 3. Figure 1 shows the relation to the popularity in associated dictionary. The percentages of extracted hyponyms in each group are shown in the graph. In general, the more popular the hyponym is in the associative concept dictionary, the more likely the word will be extracted from the corpus. Almost all (in average 94%) hyponyms associated by more than 20% of subjects are extracted from the corpus, which is much higher than the average of all groups of 67%, and hyponyms associated by less than 20% of subjects are extracted in less than 50% of the cases.
We found that there is a correlation between the frequency of the extracted hyponyms and the popularity of the associated hyponyms. For example, 20 out of 23 extracted hyponyms for "*yasai* (vegetable)" whose frequencies are more than 10 are also associated. Also, 10 out of 10 hyponyms for "vegetable" associated by more than 10 subjects (20%) appears in the corpus more than 40 times.
Some extracted hyponyms of "*doubutsu* (animal)" whose frequency is high are not associated. If a hypernym has many popular hyponyms, subjects stop associating at some point. But corpus would never get tired. "Raccoon dog" which is a very popular animal in Japan, was extracted but no one associated it.

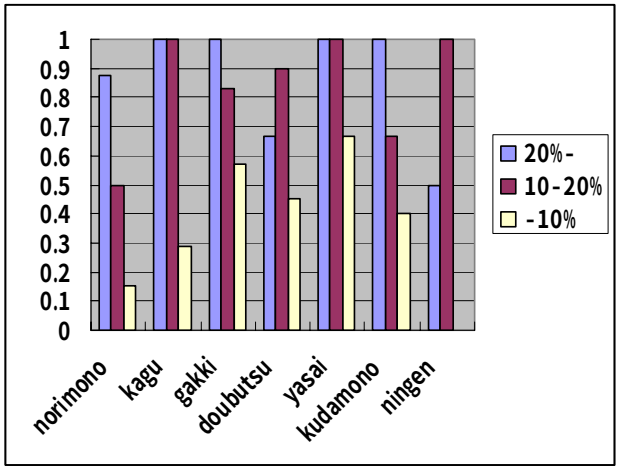

Figure 1: Relation to the popularity in associated dictionary

Conclusion

We reported on automatic hyponym extraction from a Japanese newspaper corpus. As we found some additional hyponyms over associated hyponyms, we believe the result of the automatic extraction could be a supplement to the association dictionary. For the future work, in order to improve the precision, (Cederberg and Widdows 2003) proposed to use LSA. Although their method is proposed for English, we believe it will work for Japanese and we would like to try a similar method. We are also planning to apply the method to different kinds of corpora, including a Web or patent corpus. Finally, we would like to try to extract other kinds of relationships, such as part-of and so on.

References

EDR homepage: http://www2.crl.go.jp/kk/e416/EDR
Marti A. Hearst. (1998). WordNet:An Electronic Lexical Database, Chapter 5, Automated Discovery of WordNet Relations. The MIT Press, Cambridge, Massachusetts.
Marti A. Hearst. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In Proc. of the Fourteenth International Conference on Computational Linguistics COLING'92.
Nihongo Goitaikei. (1997) Iwanami Shoten.
Okamoto, Jun. and Ishizaki, Shun. (2001). Associative concept dictionary construction and its comparison

with electronic concept dictionary. Pacific association for computational linguistics.

Satoshi Sekine, Sofia Ananiadou, Jeremy Carroll, Jun-ichi Tsujii. (1992). "Linguistic Knowledge Generator", In Proc. of the Fourteenth International Conference on Computational Linguistics COLING'92.

Scott Cederberg and Dominic Widdows. (2003). Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In Proc of the Conf. on Natural Language Learning CoNLL-2003, Edmonton, Canada.

pattern	***norimono*** **(vehicle)**	***basu*** **(bus)**	***doubutsu*** **(animal)**	***inu*** **(dog)**	***ningen*** **(human)**	***imi*** **(meaning)**	***ryokou*** **(travel)**
p1	28	11	360	33	430	37	36
	(22)	(11)	(178)	(33)	(359)	(36)	(35)
	93%	9%	80%	17%	3%	0%	3%
p2	73	20	711	10	42	150	28
	(43)	(18)	(195)	(10)	(39)	(141)	(26)
	73%	15%	97%	70%	17%	0%	4%
p3	1	-	23	3	3	2	-
	(1)		(16)	(3)	(3)	(2)	
	100%	-	96%	0%	0%	0%	-
p4	13	4	66	12	234	37	6
	(11)	(4)	(37)	(11)	(124)	(29)	(5)
	77%	25%	80%	17%	67%	0%	0%
p5	20	7	93	35	480	3958	3
	(16)	(7)	(55)	(30)	(312)	(2576)	(3)
	70%	0%	87%	77%	57%	0%	0%
p6	-	-	7	2	10	-	-
			(7)	(2)	(10)		
	-	-	71%	50%	50%	-	-
p7	2	3	100	7	114	8	7
	(2)	(3)	(31)	(6)	(68)	(8)	(6)
	100%	100%	100%	43%	100%	0%	0%
Total freq.	137	45	1360	102	1313	4192	80
Ave. precision	80%	18%	89%	44%	48%	0%	3%

Table 3: Frequency, type frequency and precision for each pattern and some hypernyms

hypernym	***norimono*** **(vehicle etc)**	***kagu*** **(furniture)**	***gakki*** **(musical instrument)**	***doubutsu*** **(animal)**	***yasai*** **(vegetable)**	***kudamono*** **(fruit)**	***ningen*** **(human)**
Coverage of associated hypo.	56% (14 / 25)	66% (10 / 15)	91% (42 / 46)	68% (26 / 38)	81% (29 / 36)	73% (22 / 30)	24% (5 / 21)
Coverage of extracted hypo.	25% (14 / 55)	14% (10 / 72)	33% (42 / 127)	10% (26 / 261)	15% (29 / 195)	29% (22 / 76)	2% (5 / 313)
Overlap hyponyms (frequency)	car (2), airplane (6), train (9)	chair (20), chest of drawers (29), desk (16)	piano (58), guitar (30), violin (19)	human being (70), dog (76), bird (23)	carrot (120), tomato (95), cabbage (86)	apple (55), orange (32), grape (31)	Japanese(2), woman(2), old man(1)
Not extracted hyponyms	motorcycle tricycle horse	low dining table, shoebox, dressing table	wood-wind instrument, recorder, tuba	amphibian, flesh-eating animal, herbivorous animal	greenish yellow vegetable, salad, organic vegetable	litch, plum, pomegranate	child, adult, man
Not associated hyponyms (frequency)	go-cart(3), tank(2), cable car(1)	air conditioner (7), dining set (2), cabinet (1)	Ocarina(5), marimba(3), handbell(2)	raccoon dog(28), penguin(7), seal(4)	parsley(9), watermelon (6), green peas(2)	blueberry(3), muscat(2), apricot(1)	I(122), myself(95), *boku*(another expression of "I")(25)

Table 4: Overlap between extracted hyponyms and associated hyponyms