Construction of Japanese Nominal Semantic Dictionary using "A NO B" Phrases in Corpora

Sadao Kurohashi, Masaki Murata[*] Yasunori Yata[†]
Mitsunobu Shimada[†] and Makoto Nagao
Graduate School of Infomatics, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto 606-8501, Japan
kuro@i.kyoto-u.ac.jp

Abstract

This paper describes a method of constructing Japanese nominal semantic dictionary, which is indispensable for text analysis, especially for indirect anaphora resolution. The main idea is to use noun phrases of "A NO(postposition) B" in corpora. Two nouns A and B in "A NO B" can have several semantic relations. By collecting "A NO B" phrases form corpora, analyzing their semantic relations, and arranging them for each "B" and each semantic relation, we can obtain a nominal semantic dictionary. The dictionary we constructed from 130M characters corpora by this method has 22,252 entries, which can be considered as a practically useful coverage. Our method for analyzing "A NO B" phrase is also original which uses a thesaurus as an attribute for decision tree.

1 Introduction

The role of dictionary is undoubtedly important in Natural Language Processing (NLP).

So far, research in NLP has mainly concerned the analysis of individual sentences. The analysis of a sentence is to clarify which element in it has relation with which by what relation. To do such an analysis, a verbal semantic dictionary, in other words, case frame dictionary is necessary. A case frame dictionary describes what kind of cases each verb has and what kinds of noun can fill a case slot. Condition on case slots can be expressed by semantic markers and/or example nouns. For example, a case frame for the verb "YOMU(read)" can be as follows:

YOMU(read)
  agent : human beings, like KARE(he), KEN(ken), SENSEI(teacher)
  object : something to be read, like HON(book), SHOSETSU(novel)

Such dictionaries with a practically useful coverage have been compiled in many institutes, mainly by hand, and used in many NLP systems (EDR, 1993; NTT, 1997).

These days, the main target of NLP has been shifting from individual sentences to a series of sentences, that is, a text. Human beings use language to communicate, and the unit of communication is not a sentence, but a text in most cases, especially in the case of written language. The NLP system can only catch enough information when it handles a text as a whole.

Similar to sentence analysis, the main part of text analysis is to clarify the relation among its constituents:

- discourse relation between text segments (cause-effect, elaboration, exemplification, etc.),
- maintaining and changing of topics,
- recovery of omission,
- direct anaphora resolution,
- indirect anaphora resolution.

To do such analyses, not only a verbal semantic dictionary, but also many other types of knowledge have to be employed, one of which is a *nominal semantic dictionary*. Similar to a verbal semantic dictionary, a nominal semantic dictionary describes what kind of nouns have what relation with each noun obligately (like obligate cases of a verb) as follows:

[*] Now at Communications Research Laboratory. E-mail: murata@crl.go.jp

[†] Now at Sharp Corporation.

KAKAKU(price)

- an attribute of something like KURUMA(car), PASOKON(personal computer), RINGO(apple), NIKU(meat)

YANE(roof)

- a part of a building like IE(house), KOYA (hut)

SENSEI(teacher)

- belongs to some institute like SHOGAKKO (elementary school), KOUKOU(high school), and
- teaches something like SUGAKU (mathematics), ONGAKU(music)

A nominal semantic dictionary is necessary for indirect anaphora resolution. Indirect anaphora is not an special, exceptional phenomena in texts, but very often used and it is very important to handle it properly for text understanding. A typical example of indirect anaphora is as follows:

> XXX announced the release of a new lap-top computer. The price is $800.

In order to find the relation between "the price" and "a new lap-top computer", a nominal semantic information about "price" has to be employed.

It is, however, almost impossible to compile a nominal semantic dictionary automatically. Then, the qustion is how we can construct a dictionary semi-automatically, or support the human compilation sufficiently. Comparing with the case of verbal dictionary, the number of noun is very big. Furthermore, if technical terms should be included, they become unlimitedly large.

Since case elements of a verb appear by the verb in a sentence, we can collect possible case elemtns of a verb by a simple parsing, or just by detecting adjoining noun and verb. On the other hand, since an anaphor and its anchor of indirect anaphora appear far away, it is almost impossible to collect them by a simple method automatically.

This paper presents how to solve this problem, namely, how to construct a nominal semantic dictionary semi-automatically.

2 Use of "A NO B" Phrases in Corpora

In Japanese, two nouns, A and B, in a phrase "A NO(postposition) B" have several semantic relations. Some relations among them can be a slot of a nominal semantic dictionary.

For example, "price" is the price of something, and in Japanese corpora we can find several phrases like "KURUMA(car) NO KAKAKU(price)" and "RINGO(apple) NO KAKAKU(price)". That is, we can obtain useful data for the entry "B" in a nominal semantic dictionay only by collecting phrases of "A NO B" from corpora.

However, all phrases of "A NO B" are not useful. For example, even if "MEIKA(maker) NO KAKAKU(price)" exists in corpora, "MEIKA(maker)" is not traded at some price, in normal cases. In other case, the phrase "WATASHI(I) NO HON(book)" does not necessarily indicate that the noun "HON(book)" has obiligate relation with "WATASHI(I)" or human beings.

Furthermore, when a phrase "A NO B" is a proper data for a nominal dictionary, it is desirable to place "A" to a proper slot of the entry "B".

These classification can be realized by the semantic analysis of "A NO B" described in the next section.

3 Semantic Analysis of "A NO B"

Japanese noun phrase "A NO B" can have one of many semantic relations listed in Table 1.

The semantic analysis of "A NO B" has been a hard problem in Japanese NLP. For this problem, Sumita et al. proposed an example-based analysis method (Sumita et al., 1990):

1. Collect many example phrases of "A NO B",
2. Give proper semantic relation to each example by hand,
3. Given an input, detect the most similar example to the input,
4. Assign the relation given to the most similar example to the input.

This is the first work that implemented an example-based method in NLP, being much

Table 1: Semantic relation of "A NO B".

1. possession (in a wide sense)	
possession	ex. WATASHI(I) NO HON(book)
whole-part*	ex. KURUMA(car) NO ENJIN(engine)
belong*	ex. HOTEL(hotel) NO JYUGYOIN(employee)
relatives*	ex. KEN(Ken) NO ANE(sister)
product/produce	ex. NIHON(Japan) NO KOME(rice)
attribute*	ex. KURUMA(car) NO KAKAKU(price)
2. A modifies B	
A:nature	ex. TANPATU(short hair) NO JYOSEI(lady)
A:action/B:agent	ex. SANPO(walk) NO HITO(man)
A:action/B:object	ex. YUNYU(import) NO RINGO(apple)
A:action/B:place*	ex. SOTSUGYOUSHIKI(graduation ceremony) NO KAIJYO(place)
A:action/B:time*	ex. SOTSUGYOU(graduation) NO JIKI(time)
A:action/B:method	ex. TSUKIN(travel to work) NO SHUDAN(way)
A:cause/B:effect*	ex. JISHIN(earthquake) NO HIGAI(damage)
A:effect/B:cause*	ex. JISHIN(earthquake) NO GENIN(cause)
A:object/B:agent*	ex. SUUGAKU(mathematics) NO SENSEI(teacher)
A:field*	ex. BENGOSHI(lawyer) NO SHIKAKU(qualification)
3. B is action	
A:agent/B:action	ex. KAZOKU(family) NO SHOUDAKU(approval)
A:object/B:action*	ex. KURUMA(car) NO HANBAI(sale)
A:goal/B:action*	ex. KYOTO(Kyoto) NO TOUCHAKU(arrival)
A:place/B:action	ex. OKUGAI(outdoor) NO ASOBI(play)
A:time/B:action	ex. 5JI(5 o'clock) NO HEITEN(close)
A:method/B:action	ex. DENSHA(train) NO TSUKIN(travel to work)
4. A is place/time	
A:place	ex. 20SEIKI(20th centry) NO ASIA(Asia)
A:time	ex. KYODAI(Kyoto University) NO TOKEIDAI(clock tower)
5. Exceptions	
idiomatic phrase	ex. CHA-NO-MA(living room)
fraction	ex. 3BUN-NO-2(two-third)
...	

more robust and easy to maintain than the conventional rule-based NLP.

The problem of example-based NLP is how to define the similarity between an input and an example. Sumita et al. caluculated the similarity between an input "A_i NO B_i" and an example "A_e NO B_e" as follows:

$$w_A \cdot sim(\mathrm{A}_i, \mathrm{A}_e) + w_B \cdot sim(\mathrm{B}_i, \mathrm{B}_e),$$

where $sim(\mathrm{A}_i, \mathrm{A}_e)$ is the similarity between A_i and A_e calculated based on the distance of the two words in a thesaurus tree, $sim(\mathrm{B}_i, \mathrm{B}_e)$ is the same for B_i and B_e, w_A and w_B are weights showing which similarity should be considered more relevant, A_i and A_e or B_i and B_e.

Such a way of caluculating words' similarity and combining them has been widely used by many researchers. However, it only has some qualitative ground, but no quantitative one.

For such a problem, Jiri and Nagao proposed a method using a thesaurus as an attribute for decision tree, being able to optimize the system on training set of examples (Jiri and Nagao, 1997). Although their method treated PP attachment ambiguity, it can be applicable to the analysis of "A NO B".

Let us explain the decision tree method for "A NO B" analysis here.

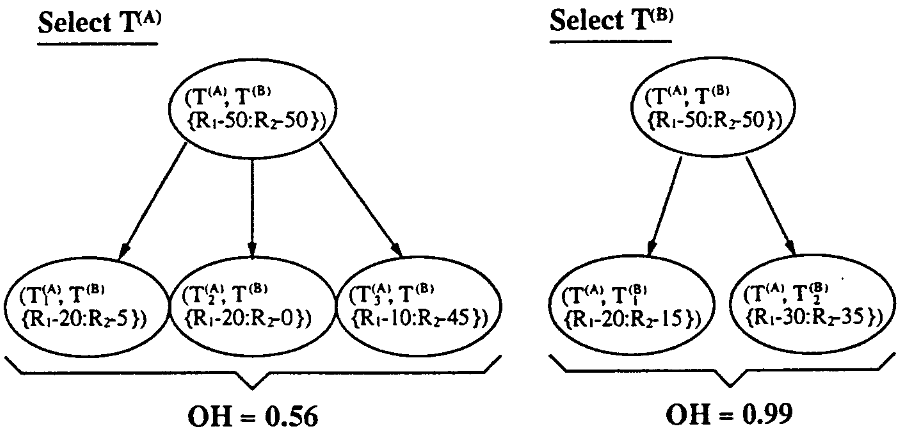
Select T(A)
Select T(B)
(T(A), T(B) {R1-50:R2-50})
(T1(A), T(B) {R1-20:R2-5})
(T2(A), T(B) {R1-20:R2-0})
(T3(A), T(B) {R1-10:R2-45})
OH = 0.56
(T(A), T(B) {R1-50:R2-50})
(T(A), T1(B) {R1-20:R2-15})
(T(A), T2(B) {R1-30:R2-35})
OH = 0.99

Figure 1: Selecting an attribute for the decision tree expansion.

Decision Tree Induction

Each example phrase "A NO B" is expressed by a triple (T_A, T_B, R_j), where T_A and T_B are the position (node) in a thesaurus matching the word A and B, respectively, R_j is the semantic relation of the phrase given by hand.

Each node in the decision tree, D, corresponds to the information expressed by a triple ($T^{(A)}$, $T^{(B)}$, S), where $T^{(A)}$ and $T^{(B)}$ are the position (node) in A-side thesaurus and B-side thesaurus, respectively, S is a subset of example phrases.

At first, the root node of the decision tree, D_{root}, corresponds to the triple ($T^{(A)}_{root}$, $T^{(B)}_{root}$, S_{all}), and it is given to the step 1 below.

1. Suppose the given decision tree node, D, corresponds to the triple ($T^{(A)}$, T^B, S). If the percentage of the major relation in the set S is greater than a threshold value (90% in our experiment), that is, S is homogenous enough, stop expanding D (D becomes a leaf of the decision tree), and the major relation of the set S is given to D. Otherwise, go to step 2.

2. Select the more informative attribute, $T^{(A)}$ or $T^{(B)}$. We consider the more informative attribute to be the one which split the node D to more homogenous subnodes. when we try $T^{(A)}$, we split the node into subnodes, each associated with a child node of $T^{(A)}$, $T^{(A)}_i$, and containing a set (S_i) of examples whose T_A is a descendant of $T^{(A)}_i$. Then, we calculate the following formula, which shows a kind of overall heterogeneity of the resulting subnodes:

$$OH = \sum_i \frac{N_i}{N} \cdot (- \sum_j \frac{N_{ij}}{N_i} \log_2 \frac{N_{ij}}{N_i}),$$

   where N is the number of examples in S, N_i is the number of examples in S_i, and N_{ij} is the number of examples in S_i which is given the j-th semantic relation. We also calculate the OH value for $T^{(B)}$, and we select one with the lower OH.

3. For the selected attribute, make subnodes as in step 2, and call the same algorithm on each subnode.

Figure 1 shows a simplified example of the above algorithm. Suppose "A NO B" phrases can be classified into only two relations, R_1 and R_2, and a given decision tree node contains 50 R_1-examples and 50 R_2-examples. If we select $T^{(A)}$ (suppose $T^{(A)}$ has three child nodes), we obtain the OH, 0.56, as in the left hand side of Figure 1; if we select $T^{(B)}$ (suppose $T^{(B)}$ has two child nodes), we obtain the OH, 0.99, as in the right hand side of Figure 1. Consequently, we select $T^{(A)}$, and split the node D into three subnodes. In the next step, the first and third nodes are analyzed by the same algorithm; the second node is not expanded any more (it is homogeneous enough).

Classification

Classification algorithm of an unseen phrase, "A_i NO B_i", using the induced decision tree is very simple. A path is traversed in the decision tree, starting at its root. At each internal node D, we follow the branch depending on the D's selected attribute ($T^{(A)}$ or $T^{(B)}$) and the thesaurus position of the input nouns (T_{A_i} or T_{B_i}). When the path reaches to a leaf of the decision tree, the phrase is assigned the majority relation of the leaf. When we cannot follow any branch at any decision tree node, D, the phrase is assigned the majority relation of D. (For example, when T_{A_i} is relatively high in the thesaurus, and at some point the decision tree tries to expand the node.)

Experiment

We did an experiment to see how well the above method works. As a thesaurus, we used EDR Concept Dictionary (EDR, 1993). We collected about 20,000 example phrases of "A NO B" from several corpora, and gave one of the semantic relations listed in Table 1 by hand. Then, we did experiments on twelve different test sets: each time, we partitioned the whole example into a training set of 19,500 phrases and a test set of 500 phrases, made a decision tree using the training set, analyzed the test set, and compared the result with the original relation given by hand. The average accuracy of the analysis was about 80%.

4 Construction of Nominal Semantic Dictionary

Our proposed method of constructing a nominal semantic dictionary is as follows: [1]

1. Collect phrases of "A NO B" from corpora, excluding syntactically ambiguous phrases like "A NO B NO C" and "A NO B C" (in both cases, "A" may modify "C", not "B").

2. For each phrase "A NO B", decide the semantic relation using the decision tree algorithm described in the previous section.

3. All examples are arranged for each "B" and each relation. "B" becomes an entry word of our dictionary, and each entry is classified by semantic relations. If the relation is not among relations marked '*' in Table 1, it is discarded from the entry [2].

In our experiments, we used Mainichi Newspaper Articles in 1995 (60M characters), and Heibonsha's World Encyclopedia (70M characters) as corpora.

From these 130M characters corpora, we collected about 620,000 types of "A NO B". Then, we analyzed these phrases, and the resulting dictionary consists of 22,252 entries, each entry has 1.5 slots on average, and each slot has 6.6 words on overage (it means that each entry has 9.9 words ($= 1.5 \times 6.6$) on average). We can say that the resulting dictionary has a practically useful wide-coverage.

Table 2 shows a couple of entries of the dictionary. We could find an interesting feature in our corpus-based dictionary. In the entry of the word "UDE (arm)", human lexicographers would make a slot of part-whole relation with "KARADA (body)" at first. In the automatic constructed dictionary, however, the major slot is *field*, with examples of "TENIS (tennis)", "SHODOU (calligraphy)". This reflects the fact that a metaphoric usage of "UDE (arm)" meaning ability or skill is much more frequent than the literal usage in real corpora. Such an adaptability to the real usage of words is an advantage of a corpus-based dictionary.

The remaining problem is how to clean up the dictionary. As mentioned in the previous section, the accuracy of the semantic analysis of "A NO B" is about 80%, resulting in many inappropriate words in the dictionary slots. For example, the entry of "KAKAKU(price)" in Table 2, *attribute* slot includes "URITE(seller)" and "KAITE(buyer)". These are the results of incorrect analysis of "URITE(seller) NO KAKAKU(price)" and "KAITE(buyer) NO KAKAKU(price)". One way of cleaning up is to introduce some machine learning method, aiming at more automatic process. However, the current dictionary is not so bad, and it's not so

[1] For an action noun, like "sale", "arrival", it is possible to utilize a verbal semantic information about its verbal form ("sell" and "arrive"). However, in this paper, we limit the discussion to the method only using "A NO B" phrases.

[2] Currently we consider semantic relations marked '*' in Table 1 can be a relation between an anaphor and the anchor. However, more investigation is necessary for this criteria.

Table 2: Example entries in the automatic constructed nominal dictionary.

KAKAKU(price)	
attribute :	RINGO(apple), BUTANIKU(port), KIN(gold), URITE(seller), MEMORI(memory), KAITE(buyer), KURUMA(car) ···
SENSEI(teacher)	
agent-object :	BIJYUTU(art), GOLF(golf), ONGAKU(music) ···
belong :	KOUKOU(high school), SHOUGAKKOU(elementary school), JYUKU(crammer) ···
YANE(roof)	
part-whole :	JYUTAKU(house), KURUMA(car), STADIUM(stadium), KOYA(hut) ···
UDE(arm)	
field :	TENNIS(tennis), SHODOU(calligraphy), KARATE(karate), ENSOU(musical performance) ···
part-whole :	IHUKU(clothes), NINGYOU(doll)

hard to clean up it by hand.

5 Conclusion

In this paper, we described a method of constructing Japanese nominal semantic dictionary using noun phrases of "A NO B" in corpora. The resulting dictionary we constructed from 130M characters corpora by this method has 22,252 entries, which can be considered as a practically useful coverage.

What we have to do next is to clean up the dictionary, since the automatic analysis of "A NO B" phrase has some errors. Another target is to employ the resulting dictionary in our text analysis system which handles direct anaphora, indirect anaphora, and omission simultaneously.

References

Japan Electronic Dictionary Research Institute Ltd. 1993. *EDR Electronic Dictionary Specifications Guide.*

Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H,, Ogura, K., Oyama, Y. and Hayashi, Y. 1997. *Japanese Lexicon.* Iwanami Publishing.

Sumita, E., Iida, H. and Kohyama, H. 1990. Translating with Examples: A New Approach to Machine Translation. *Proc. of 3rd TMI.*

Jiri Stetina and Makoto Nagao. 1997. Corpus Based PP-Attachment Ambiguity Resolution with a Semantic Dictionary. *Proc. 5th Workshop on Very Large Corpora*, Hongkong.