Japanese Case Frame Construction by Coupling the Verb and its Closest Case Component

Daisuke Kawahara
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
kawahara@pine.kuee.kyoto-u.ac.jp

Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
kuro@pine.kuee.kyoto-u.ac.jp

ABSTRACT

This paper describes a method to construct a case frame dictionary automatically from a raw corpus. The main problem is how to handle the diversity of verb usages. We collect predicate-argument examples, which are distinguished by the verb and its closest case component in order to deal with verb usages, from parsed results of a corpus. Since these couples multiply to millions of combinations, it is difficult to make a wide-coverage case frame dictionary from a small corpus like an analyzed corpus. We, however, use a raw corpus, so that this problem can be addressed. Furthermore, we cluster and merge predicate-argument examples which does not have different usages but belong to different case frames because of different closest case components. We also report on an experimental result of case structure analysis using the constructed case frame dictionary.

1. INTRODUCTION

Syntactic analysis or parsing has been a main objective in Natural Language Processing. In case of Japanese, however, syntactic analysis cannot clarify relations between words in sentences because of several troublesome characteristics of Japanese such as scrambling, omission of case components, and disappearance of case markers. Therefore, in Japanese sentence analysis, case structure analysis is an important issue, and a case frame dictionary is necessary for the analysis.

Some research institutes have constructed Japanese case frame dictionaries manually [2, 3]. However, it is quite expensive, or almost impossible to construct a wide-coverage case frame dictionary by hand.

Others have tried to construct a case frame dictionary automatically from analyzed corpora. However, existing syntactically analyzed corpora are too small to learn a dictionary, since case frame information consists of relations between nouns and verbs, which multiplies to millions of combinations. Based on such a consideration, we took the unsupervised learning strategy to Japanese case frame construction[1].

To construct a case frame dictionary from a raw corpus, we parse a raw corpus first, but parse errors are problematic in this case. However, if we use only reliable modifier-head relations to construct a case frame dictionary, this problem can be addressed. Verb sense ambiguity is rather problematic. Since verbs can have different cases and case components depending on their meanings, verbs which have different meanings should have different case frames. To deal with this problem, we collect predicate-argument examples, which are distinguished by the verb and its closest case component, and cluster them. That is, examples are not distinguished by verbs such as *naru* ‘make, become’ and *tsumu* ‘load, accumulate’, but by couples such as *tomodachi ni naru* ‘make a friend’, *byouki ni naru* ‘become sick’,*nimotsu wo tsumu* ‘load baggage’, and *keiken wo tsumu* ‘accumulate experience’. Since these couples multiply to millions of combinations, it is difficult to make a wide-coverage case frame dictionary from a small corpus like an analyzed corpus. We, however, use a raw corpus, so that this problem can be addressed. The clustering process is to merge examples which does not have different usages but belong to different case frames because of different closest case components.

2. VARIOUS METHODS FOR CASE FRAME CONSTRUCTION

We employ the following procedure of case frame construction from raw corpus (Figure 1):

1. A large raw corpus is parsed by KNP [5], and reliable modifier-head relations are extracted from the parse results. We call these modifier-head relations **examples**.

2. The extracted examples are distinguished by the verb and its closest case component. We call these data **example patterns**.

3. The example patterns are clustered based on a thesaurus. We call the output of this process **example case frames**, which is the final result of the system. We call words which compose case components **case examples**, and a group of case examples **case example group**. In Figure 1, *nimotsu* ‘baggage’, *busshi*

[1] In English, several unsupervised methods have been proposed[7, 1]. However, it is different from those that combinations of nouns and verbs must be collected in Japanese.

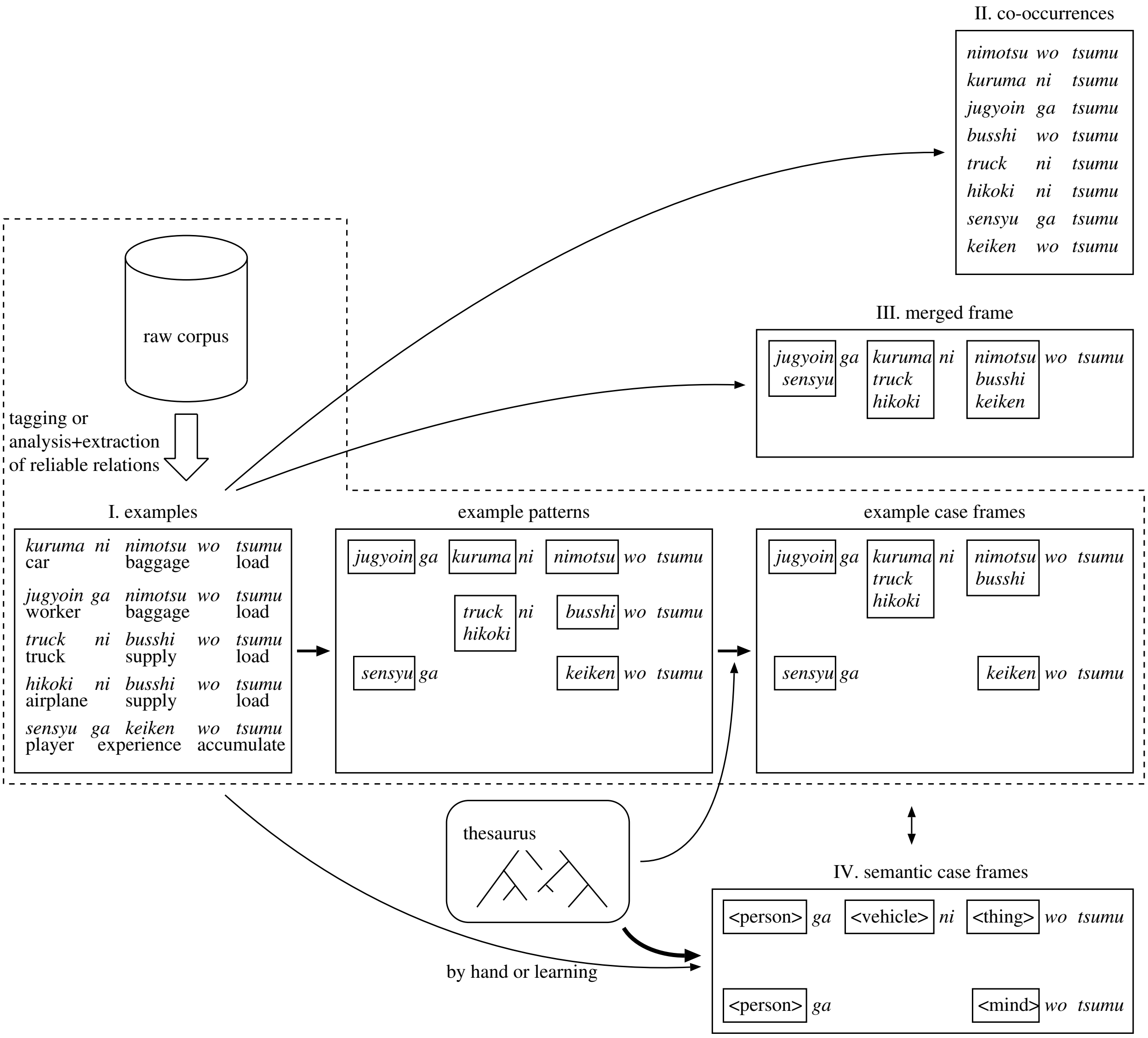

Figure 1: Several methods for case frame construction.

'supply', and *keiken* 'experience' are case examples, and {*nimotsu* 'baggage', *busshi* 'supply'} (of *wo* case marker in the first example case frame of *tsumu* 'load, accumulate') is a case example group. A **case component** therefore consists of a case example and a case marker (CM).

Let us now discuss several methods of case frame construction as shown in Figure 1.

First, examples (I of Figure 1) can be used individually, but this method cannot solve the sparse data problem. For example,

(1) *kuruma ni* *nimotsu wo* *tsumu*
car dat-CM baggage acc-CM load
(load baggage onto the car)

(2) *truck ni* *busshi wo* *tsumu*
truck dat-CM supply acc-CM load
(load supply onto the truck)

even if these two examples occur in a corpus, it cannot be judged whether the expression "*kuruma ni busshi wo tsumu*" (load supply onto the car) is allowed or not.

Secondly, examples can be decomposed into binomial relations (II of Figure 1). These co-occurrences are utilized by statistical parsers, and can address the sparse data problem. In this case, however, verb sense ambiguity becomes a serious problem. For example,

(3) *kuruma ni* *nimotsu wo* *tsumu*
car dat-CM baggage acc-CM load
(load baggage onto the car)

(4) *keiken wo* *tsumu*
experience acc-CM accumulate
(accumulate experience)

from these two examples, three co-occurrences ("*kuruma ni* *tsumu*", "*nimotsu wo tsumu*", and "*keiken wo tsumu*") are extracted. They, however, allow the incorrect expression "*kuruma ni keiken wo tsumu*" (load experience onto the car, accumulate experience onto the car).

Thirdly, examples can be simply merged into one frame (III of Figure 1). However, information quantity of this is equivalent to that of the co-occurrences (II of Figure 1), so verb sense ambiguity becomes a problem as well.

We distinguish examples by the verb and its closest case component. Our method can address the two problems above: verb sense ambiguity and sparse data.

On the other hand, semantic markers can be used as case components instead of case examples. These we call **semantic case frames** (IV of Figure 1). Constructing semantic case frames by hand leads to the problem mentioned in Section 1. Utsuro et al. constructed semantic case frames from a corpus [8]. There are three main differences to our approach: they use an annotated corpus, depend deeply on a thesaurus, and did not resolve verb sense ambiguity.

3. COLLECTING EXAMPLES

This section explains how to collect examples shown in Figure 1. In order to improve the quality of collected examples, reliable modifier-head relations are extracted from the parsed corpus.

3.1 Conditions of case components

When examples are collected, case markers, case examples, and case components must satisfy the following conditions.

Conditions of case markers

Case components which have the following case markers (CMs) are collected: *ga* (nominative), *wo* (accusative), *ni* (dative), *to* (with, that), *de* (optional), *kara* (from), *yori* (from), *he* (to), and *made* (to). We also handle **compound case markers** such as *ni-tsuite* 'in terms of', *wo-megutte* 'concerning', and others.

In addition to these cases, we introduce **time case marker**. Case components which belong to the class <time>(see below) and contain a *ni*, *kara*, or *made* CM are merged into time CM. This is because it is important whether a verb deeply relates to time or not, but not to distinguish between surface CMs.

Generalization of case examples

Case examples which have definite meanings are generalized. We introduce the following three classes, and use these classes instead of words as case examples.

<time>

- nouns which mean time
  e.g. *asa* 'morning', *haru* 'spring',
  rainen 'next year'
- case examples which contain a unit of time
  e.g. 1999*nen* 'year', 12*gatsu* 'month',
  9*ji* 'o'clock'
- words which are followed by the suffix *mae* 'before', *tyu* 'during', or *go* 'after' and do not have the semantic marker <place> on the thesaurus
  e.g. *kaku mae* 'before ··· write',
  kaigi go 'after the meeting'

<quantity>

- numerals
  e.g. *ichi* 'one', *ni* 'two', *juu* 'ten'
- numerals followed by a numeral classifier[2] such as *tsu*, *ko*, and *nin*.

  They are expressed with pairs of the class <quantity> and a numeral classifier: <quantity>*tsu*, <quantity>*ko*, and <quantity>*nin*.

  e.g. 1*tsu* → <quantity>*tsu*
  2*ko* → <quantity>*ko*

<clause>

- quotations ("··· *to*" 'that ···') and expressions which function as quotations ("··· *koto wo*" 'that ···').

  e.g. *kaku to* 'that ··· write',
  kaita koto wo 'that ··· wrote'

Exclusion of ambiguous case components

We do not use the following case components:

- Since case components which contain topic markers (TMs) and clausal modifiers do not have surface case markers, we do not use them. For example,

sono giin	*wa*	··· *wo*	*teian-shita.*
the assemblyman	TM	acc-CM	proposed

  wa is a topic marker and *giin wa* 'assemblyman TM' depends on *teian-shita* 'proposed', but there is no case marker for *giin* 'assemblyman' in relation to *teian-shita* 'proposed'.

··· *wo*	*teian-shiteiru*	*giin ga* ···
acc-CM	proposing	assemblyman

  "··· *wo teian-shiteiru*" is a clausal modifier and *teian-shiteiru* 'proposing' depends on *giin* 'assemblyman', but there is no case marker for *giin* 'assemblyman' in relation to *teian-shiteiru* 'proposing'.

- Case components which contain a *ni* or *de* case marker are sometimes used adverbially. Since they have the optional relation to their verbs, we do not use them.

  e.g. *tame ni* 'because of',
  mujouken ni 'unconditionally',
  ue de 'in addition to'

For example,

30nichi	*ni*	*souri daijin*	*ga*
30th	on	prime minister	nom-CM

sono 2nin	*ni*
those two people	dat-CM

syou	*wo*	*okutta*
award	acc-CM	gave

[2] Most nouns must take a numeral classifier when they are quantified in Japanese. An English equivalent to it is 'piece'.

(On 30th the prime minister gave awards to those two people.)

from this sentence, the following example is acquired.

<time>:time-CM *daijin:ga*
minister:nom-CM

<quantity>*nin:ni* *syou:wo* *okuru*
people:dat-CM award acc-CM give

3.2 Conditions of verbs

We collect examples not only for verbs, but also for adjectives and noun+copulas[3]. However, when a verb is followed by a causative auxiliary or a passive auxiliary, we do not collect examples, since the case pattern is changed.

3.3 Extraction of reliable examples

When examples are extracted from automatically parsed results, the problem is that the parsed results inevitably contain errors. Then, to decrease influences of such errors, we discard modifier-head relations whose parse accuracies are low and use only reliable relations.

KNP employs the following heuristic rules to determine a head of a modifier:

HR1 KNP narrows the scope of a head by finding a clear boundary of clauses in a sentence. When there is only one candidate verb in the scope, KNP determines this verb as the head of the modifier.

HR2 Among the candidate verbs, verbs which rarely take case components are excluded.

HR3 KNP determines the head according to the preference: a modifier which is not followed by a comma depends on the nearest candidate, and a modifier with a comma depends on the second nearest candidate.

Our approach trusts HR1 but not HR2 and HR3. That is, modifier-head relations which are decided in HR1 (there is only one candidate of the head in the scope) are extracted as examples, but relations which HR2 and HR3 are applied to are not extracted. The following examples illustrate the application of these rules.

(5) *kare wa kai-tai hon wo*
he TM want to buy book acc-CM

takusan mitsuketa node,
a lot found because

Tokyo he okutta.
Tokyo to sent

(Because he found a lot of books which he wants to buy, he sent them to Tokyo.)

In this example, an example which can be extracted without ambiguity is "*Tokyo he okutta*" 'sent ϕ to Tokyo' at the end of the sentence. In addition, since *node* 'because' is analyzed as a clear boundary of clauses, the head candidate of *hon wo* 'book acc-CM' is only *mitsuketa* 'find', and this is also extracted.

Verbs excluded from head candidates by HR2 possibly become heads, so we do not use the examples which HR2 is applied to. For example, when there is a strong verb right after an adjective, this adjective tends not to be a head of a case component, so it is excluded from head candidates.

(6) *Hi no mawari ga hayaku*
fire of spread nom-CM rapidly

sukuidase-nakatta.
could not save

(The fire spread rapidly, so ϕ_1 could not save ϕ_2.)

In this example, the correct head of *mawari ga* 'spread' is *hayaku* 'rapidly'. However, since *hayaku* 'rapidly' is excluded from the head candidates, the head of *mawari ga* 'spread' is analyzed incorrectly.

We show an example of the process HR3:

(7) *kare ga shitsumon ni*
he nom-CM question acc-CM

sentou wo kitte kotaeta.
lead acc-CM take answered

(He took the lead to answer the question.)

In this example, head candidates of *shitsumon ni* 'question acc-CM' are *kitte* 'take' and *kotaeta* 'answered'. According to the preference "modify the nearer head", KNP incorrectly decides the head is *kitte* 'take'. Like this example, when there are many head candidates, the decided head is not reliable, so we do not use examples in this case.

We extracted reliable examples from Kyoto University Corpus[6], that is a syntactically analyzed corpus, and evaluated the accuracy of them. The accuracy of all the case examples which have the target cases was 90.9%, and the accuracy of the reliable examples was 97.2%. Accordingly, this process is very effective.

4. CONSTRUCTION OF EXAMPLE CASE FRAMES

As shown in Section 2, when examples whose verbs have different meanings are merged, a case frame which allows an incorrect expression is created. So, for verbs with different meanings, different case frames should be acquired.

In most cases, an important case component which decides the sense of a verb is the closest one to the verb, that is, the verb sense ambiguity can be resolved by coupling the verb and its closest case component. Accordingly, we distinguish examples by the verb and its closest case component. We call the case marker of the closest case component **closest case marker**.

The number of example patterns which one verb has is equal to that of the closest case components. That is, example patterns which have almost the same meaning are individually handled as follows:

(8) *jugyoin*:*ga* *kuruma*:*ni*
worker:nom-CM car:dat-CM

nimotsu:*wo* *tsumu*
baggage:acc-CM load

(9) {*truck,hikoki*}:*ni*
{truck,airplane}:dat-CM

busshi:*wo* *tsumu*
supply:acc-CM load

In order to merge example patterns that have almost the same meaning, we cluster example patterns. The final ex-

[3] In this paper, we use 'verb' instead of 'verb/adjective or noun+copula' for simplicity.

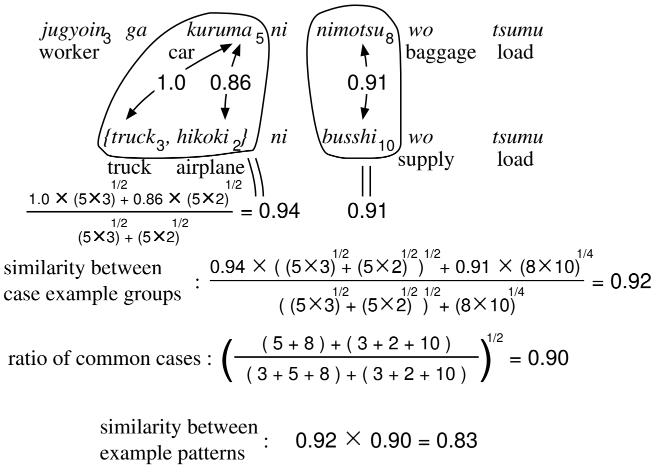
jugyoin$_3$ ga
worker
kuruma$_5$ ni
car
1.0
0.86
{truck$_3$, hikoki$_2$} ni
truck airplane
nimotsu$_8$ wo
baggage
0.91
busshi$_{10}$ wo
supply
tsumu
load
tsumu
load
$\frac{1.0 \times (5\times3)^{1/2} + 0.86 \times (5\times2)^{1/2}}{(5\times3)^{1/2} + (5\times2)^{1/2}} = 0.94$
0.91
similarity between case example groups : $\frac{0.94 \times ((5\times3)^{1/2} + (5\times2)^{1/2})^{1/2} + 0.91 \times (8\times10)^{1/4}}{((5\times3)^{1/2} + (5\times2)^{1/2})^{1/2} + (8\times10)^{1/4}} = 0.92$
ratio of common cases : $\left(\frac{(5+8)+(3+2+10)}{(3+5+8)+(3+2+10)}\right)^{1/2} = 0.90$
similarity between example patterns : $0.92 \times 0.90 = 0.83$

Figure 2: Example of calculating the similarity between example patterns (Numerals in the lower right of examples represent their frequencies.)

ample case frames consist of the example pattern clusters. The detail of the clustering is described in the following section.

4.1 Similarity between example patterns

The clustering of example patterns is performed by using the similarity between example patterns. This similarity is based on the similarities between case examples and the ratio of common cases. Figure 2 shows an example of calculating the similarity between example patterns.

First, the similarity between two examples e_1, e_2 is calculated using the NTT thesaurus as follows:

$$sim_e(e_1, e_2) = max_{x \in s_1, y \in s_2}\, sim(x, y)$$

$$sim(x, y) = \frac{2L}{l_x + l_y}$$

where x, y are semantic markers, and s_1, s_2 are sets of semantic markers of e_1, e_2 respectively[4]. l_x, l_y are the depths of x, y in the thesaurus, and the depth of their lowest (most specific) common node is L. If x and y are in the same node of the thesaurus, the similarity is 1.0, the maximum score based on this criterion.

Next, the similarity between the two case example groups E_1, E_2 is the normalized sum of the similarities of case examples as follows:

$$sim_E(E_1, E_2) = \frac{\sum_{e_1 \in E_1} \sum_{e_2 \in E_2} \sqrt{|e_1||e_2|}\, sim_e(e_1, e_2)}{\sum_{e_1 \in E_1} \sum_{e_2 \in E_2} \sqrt{|e_1||e_2|}}$$

where $|e_1|, |e_2|$ represent the frequencies of e_1, e_2 respectively.

The ratio of common cases of example patterns F_1, F_2 is calculated as follows:

$$cs = \sqrt{\frac{\sum_{i=1}^{n} |E_{1cc_i}| + \sum_{i=1}^{n} |E_{2cc_i}|}{\sum_{i=1}^{l} |E_{1c1_i}| + \sum_{i=1}^{m} |E_{2c2_i}|}}$$

where the cases of example pattern F_1 are $c1_1, c1_2, \cdots, c1_l$, the cases of example pattern F_2 are $c2_1, c2_2, \cdots, c2_m$, and the common cases of F_1 and F_2 is $cc_1, cc_2, \cdots, cc_n$. E_{1cc_i} is the case example group of cc_i in F_1. E_{2cc_i}, E_{1c1_i}, and E_{2c2_i} are defined in the same way. The square root in this equation decreases influences of the frequencies.

The similarity between F_1 and F_2 is the product of the ratio of common cases and the similarities between case example groups of common cases of F_1 and F_2 as follows:

$$score = cs \cdot \frac{\sum_{i=1}^{n} \sqrt{w_i}\, sim_E(E_{1cc_i}, E_{2cc_i})}{\sum_{i=1}^{n} \sqrt{w_i}}$$

$$w_i = \sum_{e_1 \in E_{1cc_i}} \sum_{e_2 \in E_{2cc_i}} \sqrt{|e_1|\,|e_2|}$$

where w_i is the weight of the similarities between case example groups.

4.2 Selection of semantic markers of example patterns

The similarities between example patterns are deeply influenced by semantic markers of the closest case components. So, when the closest case components have semantic ambiguities, a problem arises. For example, when clustering example patterns of *awaseru* 'join, adjust', the pair of example patterns (*te* 'hand', *kao*, 'face')[5] is created with the common semantic marker <part of an animal>, and (*te* 'method', *syouten* 'focus') is created with the common semantic marker <logic, meaning>. From these two pairs, the pair (*te* 'hand', *kao* 'face', *syouten* 'focus') is created, though <part of an animal> is not similar to <logic, meaning> at all.

To address this problem, we select one semantic marker of the closest case component of each example pattern in order of the similarity between example patterns as follows:

1. In order of the similarity of a pair, (p, q), of two example patterns, we select semantic markers of the closest case components, n_p, n_q of p, q. The selected semantic markers s_p, s_q maximize the similarity between n_p and n_q.
2. The similarities of example patterns related to p, q are recalculated.
3. These two processes are iterated while there are pairs of two example patterns, of which the similarity is higher than a threshold.

4.3 Clustering procedure

The following is the clustering procedure:

1. Elimination of example patterns which occur infrequently

   Target example patterns of the clustering are those whose closest case components occur more frequently than a threshold. We set this threshold to 5.

[4] In many cases, nouns have many semantic markers in NTT thesaurus.

[5] Example patterns are represented by the closest case components.

2. Clustering of example patterns which have the same closest CM

   (a) Similarities between pairs of two example patterns which have the same closest CM are calculated, and semantic markers of closest case components are selected. These two processes are iterated as mentioned in 4.2.

   (b) Each example pattern pair whose similarity is higher than some threshold is merged.

3. Clustering of all the example patterns

   The example patterns which are output by 2 are clustered. In this phase, it is not considered whether the closest CMs are the same or not. The following example patterns have almost the same meaning, but they are not merged by 2 because of the different closest CM. This clustering can merge these example patterns.

(10) {*busshi,kamotsu*}:*wo*
{supply,cargo}:acc-CM

truck:*ni* *tsumu*
truck:dat-CM load

(11) {*truck,hikoki*}:*ni*
{truck,airplane}:dat-CM

{*nimotsu,busshi*}:*wo* *tsumu*
{baggage,supply}:acc-CM load

5. SELECTION OF OBLIGATORY CASE MARKERS

If a CM whose frequency is lower than other CMs, it might be collected because of parsing errors, or has little relation to its verb. So, we set the threshold for the CM frequency as $2\sqrt{mf}$, where mf means the frequency of the most found CM. If the frequency of a CM is less than the threshold, it is discarded. For example, suppose the most frequent CM for a verb is *wo*, 100 times, and the frequency of *ni* CM for the verb is 16, *ni* CM is discarded (since it is less than the threshold, 20).

However, since we can say that all the verbs have *ga* (nominative) CMs, *ga* CMs are not discarded. Furthermore, if an example case frame do not have a *ga* CM, we supplement its *ga* case with semantic marker <person>.

6. CONSTRUCTED CASE FRAME DICTIONARY

We applied the above procedure to Mainichi Newspaper Corpus (9 years, 4,600,000 sentences). We set the threshold of the clustering 0.80. The criterion for setting this threshold is that case frames which have different case patterns or different meanings should not be merged into one case frame. Table1 shows examples of constructed example case frames.

From the corpus, example case frames of 71,000 verbs are constructed; the average number of example case frames of a verb is 1.9; the average number of case slots of a verb is 1.7; the average number of example nouns in a case slot is 4.3. The clustering led a decrease in the number of example case frames of 47%.

Table 1: Examples of the constructed case frames(* means the closest CM).

verb	CM	case examples
*kau*1 'buy'	*ga* *wo** *de*	person, passenger stock, land, dollar, ticket shop, station, yen
*kau*2	*ga* *wo**	treatment, welfare, postcard anger, disgust, antipathy
⋮	⋮	⋮
*yomu*1 'read'	*ga* *wo**	student, prime minister book, article, news paper
*yomu*2	*ga* *wo* *de**	<person> talk, opinion, brutality news paper, book, textbook
*yomu*3	*ga* *wo**	<person> future
⋮	⋮	⋮
*tadasu*1 'examine'	*ga* *wo** *ni tsuite*	member, assemblyman opinion, intention, policy problem, <clause>, bill
*tadasu*2 'improve'	*ga* *wo**	chairman, oneself position, form
⋮	⋮	⋮
*kokuchi*1 'inform'	*ga* *ni**	doctor the said person
*kokuchi*2	*ga* *wo** *ni**	colleague infection, cancer patient, family
*sanseida*1 'agree'	*ga* *ni**	<person> opinion, idea, argument
*sanseida*2	*ga* *ni**	<person> <clause>

As shown in Table1, example case frames of noun+copulas such as *sanseida* 'positiveness+copula (agree)', and compound case markers such as *ni-tsuite* 'in terms of' of *tadasu* 'examine' are acquired.

7. EXPERIMENTS AND DISCUSSION

Since it is hard to evaluate the dictionary statically, we use the dictionary in case structure analysis and evaluate the analysis result. We used 200 sentences of Mainichi Newspaper Corpus as a test set. We analyzed case structures of the sentences using the method proposed by [4]. As the evaluation of the case structure analysis, we checked whether cases of ambiguous case components (topic markers and clausal modifiers) are correctly detected or not. The evaluation result is presented in Table 2. The baseline is the result by assigning a vacant case in order of 'ga', 'wo', and 'ni'. When we do not consider parsing errors to evaluate the case detection, the accuracy of our method for topic markers was 96% and that for clausal modifiers was 76%. The baseline accuracy for topic markers was 91% and that for clausal modifiers was 62%. Thus we see our method is superior to the baseline.

Table 2: The accuracy of case detection.

		correct case detection	incorrect case detection	parsing error
our method	topic marker	85	4	13
	clausal modifier	48	15	2
baseline	topic marker	81	8	13
	clausal modifier	39	24	2

The following are examples of analysis results[6]:

(1) $_{1}\underline{\textit{ookurasyo}}_{\bigcirc ga}$ *wa* *ginko* *ga*
the Ministry of Finance TM bank nom-CM

$^{2}\boxed{\textit{tsumitate-teiru}}$ $_{2}\underline{\textit{ryuhokin}}_{\bigcirc wo}$ *no*
deposit reserve fund of

torikuzushi *wo* $^{3}\boxed{\textit{mitomeru}}$
consume acc-CM consent

$_{3}\underline{\textit{houshin}}_{\times ni}\dagger$ *wo* $^{1}\boxed{\textit{kimeta}}$.
policy acc-CM decide

(The Ministry of Finance decided the policy of consenting to consume the reserve fund which the banks have deposited.)

(2) *korera* *no* $_{1}\underline{\textit{gyokai}}_{\times wo}\ddagger$ *wa* *seijiteki*
these industry TM political

hatsugenryoku *ga* *tsuyoi* *toiu*
voice nom-CM strong

tokutyo *ga* $^{1}\boxed{\textit{aru}}$.
characteristic nom-CM have
(These industries have the characteristic of strong political voice.)

Analysis errors are mainly caused by two phenomena. The first is clausal modifiers which have no case relation to the modifees such as "$\cdots$ *wo mitomeru houshin*" 'policy of consenting $\cdots$' († above). The Second is verbs which take two *ga* 'nominative' case markers (one is *wa* superficially) such as "*gyokai* $\underline{\textit{wa}}$ $\cdots$ *toiu tokutyo* $\underline{\textit{ga}}$ *aru*" 'industries have the characteristic of $\cdots$' (‡ above). Handling these phenomena is an area of future work.

8. CONCLUSION

We proposed an unsupervised method to construct a case frame dictionary by coupling the verb and its closest case component. We obtained a large case frame dictionary, which consists of 71,000 verbs. Using this dictionary, we can detect ambiguous case components accurately. We plan to exploit this dictionary in anaphora resolution in the future.

9. ACKNOWLEDGMENTS

The research described in this paper was supported in part by JSPS-RFTF96P00502 (The Japan Society for the Promotion of Science, Research for the Future Program).

10. REFERENCES

[1] T. Briscoe and J. Carroll. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 356–363, 1997.

[2] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, and Y. O. Y. Hayashi, editors. *Japanese Lexicon*. Iwanami Publishing, 1997.

[3] Information-Technology Promotion Agency, Japan. *Japanese Verbs : A Guide to the IPA Lexicon of Basic Japanese Verbs*. 1987.

[4] S. Kurohashi and M. Nagao. A method of case structure analysis for japanese sentences based on examples in case frame dictionary. In *IEICE Transactions on Information and Systems*, volume E77-D No.2, 1994.

[5] S. Kurohashi and M. Nagao. A syntactic analysis method of long japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4), 1994.

[6] S. Kurohashi and M. Nagao. Building a japanese parsed corpus while improving the parsing system. In *Proceedings of The First International Conference on Language Resources & Evaluation*, pages 719–724, 1998.

[7] C. D. Manning. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31th Annual Meeting of ACL*, pages 235–242, 1993.

[8] T. Utsuro, T. Miyata, and Y. Matsumoto. Maximum entropy model learning of subcategorization preference. In *Proceedings of the 5th Workshop on Very Large Corpora*, pages 246–260, 1997.

[6] The underlined words with ◯ are correctly analyzed, but ones with × are not. The detected CMs are shown after the underlines.