Japanese Case Structure Analysis by Unsupervised Construction of a Case Frame Dictionary

Daisuke Kawahara, Nobuhiro Kaji and **Sadao Kurohashi**
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606–8501, Japan
{kawahara,kaji,kuro}@pine.kuee.kyoto-u.ac.jp

Abstract

In Japanese, case structure analysis is very important to handle several troublesome characteristics of Japanese such as scrambling, omission of case components, and disappearance of case markers. However, for lack of a wide-coverage case frame dictionary, it has been difficult to perform case structure analysis accurately. Although several methods to construct a case frame dictionary from analyzed corpora have been proposed, they cannot avoid data sparseness problem. This paper proposes an unsupervised method of constructing a case frame dictionary from an enormous raw corpus by using a robust and accurate parser. It also provides a case structure analysis method based on the constructed 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 (Ikehara et al., 1997; Information-Technology Promotion Agency, Japan, 1987). 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 (Utsuro et al., 1998). 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 following unsupervised learning strategy to the Japanese case structure analysis:

1. At first, a robust and accurate parser is developed, which does not utilize a case frame dictionary,
2. a very large corpus is parsed by the parser,
3. reliable noun-verb relations are extracted from the parse results, and a case frame dictionary is constructed from them, and
4. the dictionary is utilized for case structure analysis.

2 Characteristics of Japanese language and necessity of case structure analysis

In Japanese, postpositions function as case markers (CMs) and a verb is final in a sentence. The basic structure of a Japanese sentence is as follows:

(1) *kare ga* *coat wo* *kiru.*
he nominative-CM coat accusative-CM wear
(he wears a coat)

A clause modifier is left to the modified noun as follows:

(2) *kare ga* *kite-iru* *coat*
he nom-CM wear coat
(the coat he wears)

The modified noun followed by a postposition then becomes a case component of a matrix verb. The typical structure of a Japanese complex sentence is as follows:

(3) *boushi no iro wa kite-iru*
hat of color topic-marker wear
coat ni awaseru.
coat dative-CM harmonize
(ϕ harmonizes the color of his/her hat with the coat he/she wears)

In terms of automatic analysis, the problematic characteristics of Japanese sentences can be summarized as follows:

1. Case components are often scrambled or omitted.
2. Case-marking postpositions disappear when case components are accompanied by topic-markers or other special postpositions meaning 'just', 'also' and others.
   ex) *kare wa coat mo kite-iru.*
   he topic-marker coat also wear
   (He wears a coat also)
3. A noun modified by a clause is usually a case component for the verb of the modifying clause. However, there is no case-marker for their relation. In case of sentence 3, there is no case-marker for *coat* in relation to *kite-iru* 'wear'. Note that *ni* (dative-CM) of *coat ni* does not show the case to *kite-iru* 'wear', but to *awaseru* 'harmonize'.
4. Sentence 3 exhibits a typical structural ambiguity in a Japanese sentence. That is, *iro wa* 'color topic-marker' possibly modifies *kite-iru* 'wear' or *awaseru* 'harmonize'.

In English, sentence structure is rather rigid, and word order (the position in relation to the verb) clearly defines cases. In Japanese, however, the problem 1 above makes word order useless, and CMs constitute the only information for detecting cases.

Nevertheless, CMs often disappear because of the problems 2 and 3, which means that simple syntactic analysis cannot clarify cases sufficiently. For example, given an input sentence:

(4) *kare wa Deutsch-go mo hanasu.*
he topic-marker German also speak
(he speaks German also)

a simple syntactic analysis just detects both *kare* 'he' and *Deutsch-go* 'German' modifies *hanasu* 'speak', but tells nothing about which is subject and object. This analysis result is not sufficient for subsequent NLP applications like Japanese to English machine translation.

Then, what we need to do is a case structure analysis based on a case frame dictionary, or a subcat, of each verb as follows:

hanasu 'speak':	
ga (nom)	*kare* 'he', *hito* 'person'
wo (acc)	*eigo* 'English', *kotoba* 'language'
kiru 'wear':	
ga (nom)	*kare* 'he', *hito* 'person'
wo (acc)	*fuku* 'cloth', *coat* 'coat'
awaseru 'harmonize':	
ga (nom)	*kare* 'he', *hito* 'person'
wo (acc)	*iro* 'color'
ni (dat)	*fuku* 'cloth'

Consultation of such a dictionary can easily find that *kare* 'he' is a nominative case and *Deutsch-go* 'German' is an accusative case in the sentence 4.

Furthermore, a case frame dictionary can solve the problem 4 above, that is, some part of structural ambiguity in sentences. In case of sentence 3, a proper head for *iro wa* 'color topic-marker' can be selected by consulting case slots of *kiru* 'wear' and those of *awaseru* 'harmonize'.

3 Unsupervised construction of a case frame dictionary

This section explains how to construct a case frame dictionary from corpora automatically.

As mentioned in the introduction section, it is quite expensive, or almost impossible to construct a wide-coverage case frame dictionary by hand. In Japanese, some noun + copula works like an adjective. For example, *sansei da* 'positiveness + copula' can take *ga* case and *ni* case. However, such case frames are rarely covered by the existing handmade dictionaries[1].

Furthermore, existing handmade dictionaries cover typical obligatory cases like *ga* (nominative), *wo* (accusative), *ni* (dative), but do not cover compound case markers such as *ni-kanshite* 'in terms of', *wo-megutte* 'concerning' and others.

Then, we tried to construct an example-based case frame dictionary from corpora, which de-

[1]Our method collects case frames not only for verbs, but also for adjectives and nouns+copula. In this paper, we use 'verb' instead of 'verb/adjective or noun + copula' for simplicity.

Table 1: The accuracy of KNP.

ga nom.	*wo* acc.	*ni* dative	*kara* from	*made* to	*yori* from	*wa, mo* topic-marker	clause modifying verbs	clause modifying nouns	Total
91.2%	97.7%	94.2%	83.8%	85.3%	82.8%	88.0%	84.3%	95.5%	91.3%

scribes what kind of cases each verb has and what kind of nouns can fill a case slot. Very large syntactically analyzed corpora could be useful to construct such a dictionary. However, corpus annotation costs very much and existing analyzed corpora are too small from the view point of case frame learning. For example, in Kyoto University Corpus which consists of about 40,000 analyzed sentences of newspaper articles, very basic verbs like *tetsudau* 'help' or *uketsukeru* 'accept' appear only 10 times or 15 times respectively. It is obvious that such small data are insufficient for automatic case frame learning. That is, case frame learning must be done from enormous unanalyzed corpora, in unsupervised way[2].

3.1 Good parser

NLP research group at Kyoto University has been developing a robust and accurate parsing system, KNP, over the last ten years (Kurohashi and Nagao, 1994; Kurohashi and Nagao, 1998). This parser has the following advantages:

- Japanese is an agglutinative language, and several function words (auxiliary verbs, suffixes, and postpositions) often appear together and in many cases compositionality does not hold among them. KNP treats such function words carefully and precisely.
- KNP detects scopes of coordination structures well based on their parallelism.
- KNP employs several heuristic rules to produce unique parses for the input sentences.

The accuracy of KNP is shown in Table 1, which counted whether each phrase modifies a proper head or not. The overall accuracy was around 90%, and the accuracy concerning case components varies from 82% to 98%.

[2]In English, several unsupervised methods have been proposed (Manning, 1993; Briscoe and Carroll, 1997). However, as mentioned in Section 3, automatic Japanese case analysis is much harder than English.

We can collect pairs of verbs and case components from the automatic analyses of large corpora by KNP.

3.2 Coping with two problems

The quality of automatic case frame learning could be negatively influenced by the following two problems:

Word sense ambiguity: A verb sometimes has various usages and possibly has several case frames depending on its usages.

Structural ambiguity: KNP performs fairly well, but automatic parse results inevitably contain errors.

The following sections explain how to solve these problems.

3.2.1 Word sense ambiguity

If a verb has two or more meanings and their case frame patterns differ, we have to disambiguate the sense of each occurrence of the verb in a corpus first, and collect case components for each sense respectively. However, unsupervised word sense disambiguation of free texts is one of the most difficult problems in NLP. At the very beginning, even the definition of word senses is open to question.

To cope with this problem, we made a very simple but useful assumption: a light verb has different case frames depending on its main case component; an ordinary verb has a unique case frame even if it has two or more meanings. For example, the case frame of the verb *naru* 'become' differs depending on its *ni* (dative) case as follows:

···	*ga*	*byouki ni naru*	
	nom.	become ill	
···	*ga*	··· *to*	*tomodachi ni naru*
	nom.	with	become a friend

In most cases, the main case components are placed just in front of the light verbs so that the automatic parser can detect their relations

Table 2: Examples of the constructed case frames.

verbs	case markers	example nouns
tasukeru 'help'	*ga* (nom)	husband, person, child, staff, I, suspect, faculty, ...
	wo (acc)	job, shop, farmwork, preparation, election, move, ...
	ni (dat)	son, friend, ambassador, member, thank, holiday, ...
	de (op)	volunteer, affair, office, reward, house, headquarters, ...
yomu 'read'	*ga* (nom)	person, I, child, adult, parent, teacher, ...
	wo (acc)	newspaper, book, magazine, article, novel, letter, ...
	ni (dat)	child, person, daughter, teacher, student, reader, ...
	de (op)	newspaper, book, magazine, library, classroom, bathroom, ...

reliably. Therefore, as for five major and troublesome light verbs (*suru* 'do', *naru* 'become', *aru* 'is $\cdots$', *iu* 'say', *nai* 'not'), their case frames are distinguished depending on their left neighbouring case components. For other verbs, we assume a unique case frame.

3.2.2 Structural ambiguity

As shown in Table 1, KNP detects heads of case components in fairly high accuracy. However, in order to collect much reliable data, we discarded modifier-head relations in the automatically parsed corpora in the following cases:

- When CMs of case components disappear because of topic markers or others.
- When the verb is followed by a causative auxiliary or a passive auxiliary, the case pattern is changed and the trace in KNP is not so reliable.

Based on the conditions above, case components of each verb are collected from the parsed corpora, and the collected data are considered as case frames of verbs. However, if the frequency of a CM is very low compared to other CMs, it might be collected because of parse errors. 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).

3.3 Constructed case frame dictionary

We applied the above procedure to Mainichi Newspaper Corpus (7 years, 3,600,000 sentences). From the corpus, case frames of 23,497 verbs are constructed; the average number of case slots of a verb is 2.8; the average number of example nouns in a case slot is 33.6. Table 2 shows examples of constructed case frames.

Although the constructed data look appropriate in most cases, it is hard to evaluate a dictionary statically. In the next section, we use the dictionary in case structure analysis and evaluate the analysis result, which also implies an evaluation of the dictionary itself.

4 Case structure analysis using the constructed case frame dictionary

4.1 Matching of an input sentence and a case frame

The basic procedure in case structure analysis is to match an input sentence with a case frame, as shown in Figure 1.

The matching of case components in an input and case slots in a case frame is done on the following conditions:

1. When a case component has a CM, it must be assigned to the case slot with the same CM.
2. When a case component does not have a CM, it can be assigned to the *ga*, *wo*, or *ni* CM slot.
3. Only one case component can be assigned to a case slot (unique case assignment constraint).

The conditions above may produce multiple matching patterns, and to select the proper one among them, nouns of case components are compared with examples in case slots of the dictionary.

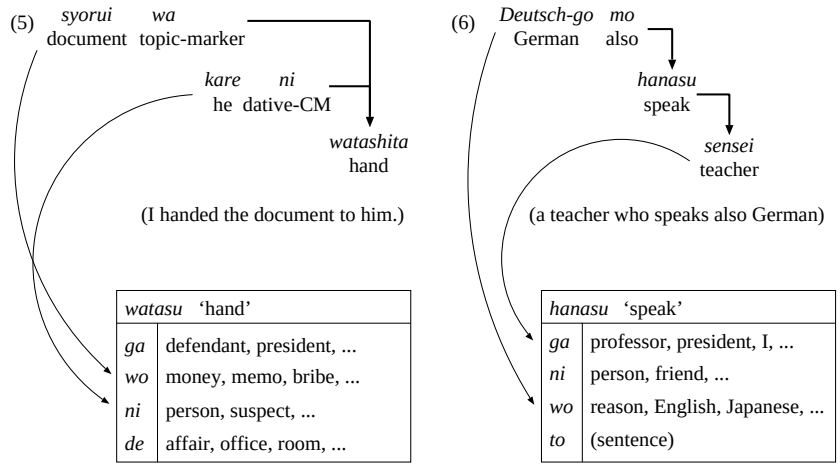

Figure 1: Matching of an input sentence and a case frame.

Even though a 3,600,000 sentences corpus was used for learning, examples in case slots are still sparse, and an input noun mostly does not match exactly an example in the dictionary. Then, a thesaurus is employed to solve this problem.

In our experiments, NTT Semantic Feature Dictionary (Ikehara et al., 1997) is employed as a thesaurus. Suppose we calculate the similarity between w_1 and w_2, their depth is d_1 and d_2 in the thesaurus, and the depth of their lowest (most specific) common node is d_c, the similarity score between them is calculated as follows:

$$sim(w_1, w_2) = (d_c \times 2)/(d_1 + d_2).$$

If w_1 and w_2 are in the same node of the thesaurus, the similarity is 1.0, the maximum score based on this criteria. If w_1 and w_2 are identical, the similarity is 1.0, of course.

The score of case assignment is the best similarity between the input noun and examples in the case slots. The score of a matching pattern is the sum of scores of case assignments in it. If two or more patterns meet the above conditions, one which has the best score is selected as a final result.

In the case of sentence 5 in Figure 1, *kare ni* 'he dative-CM' is assigned to the *ni* case slot. Then, *syorui wa* 'document topic-marker' can be assigned to the *ga* or *wo* case slot. By calculating similarity between *syorui* and *ga*-slot examples and *wo*-slot examples, it is considered to be assigned to the *wo* slot.

In case of sentence 6, none of the case components has a CM. Based on similarity calculation, *Deutsch-go* is assigned to *wo*, *sensei* is assigned to *ga*.

4.2 Parsing with case structure analysis

A complex sentence which contains a clausal modifier exhibits a typical structural ambiguity of Japanese; case components left to a verb of a clausal modifier, V_c, possibly modify V_c or a matrix verb V_m.

For example, in sentence 3, *iro wa* 'color topic-marker' possibly modifies *kite-iru* 'wear' or *awaseru* 'harmonize'.

KNP, a rule-based parser, handles this type of ambiguity as follows. If a case component is followed by a comma, it is treated as modifying V_m; if not, it is treated as modifying V_c. Although this heuristic rule usually explains real data very well, sentence 3 will be analyzed incorrectly.

Parsing which utilizes a case frame dictionary can consider which is a proper head, V_c or V_m, for an ambiguous case component by comparing examples in the case slots of V_c and V_m. Such a consideration must be done considering what other case components modify V_c and V_m, since the assigned case slot of a case component might differ depending on the candidate structure of the sentence due to the unique case assignment constraint.

Therefore, it is necessary to expand the structural ambiguity and consider all the possible structures for an input. So, we calculate the matching score of all pairs of case components and verbs in all possible structures of the sentence, and select the best structure based on the

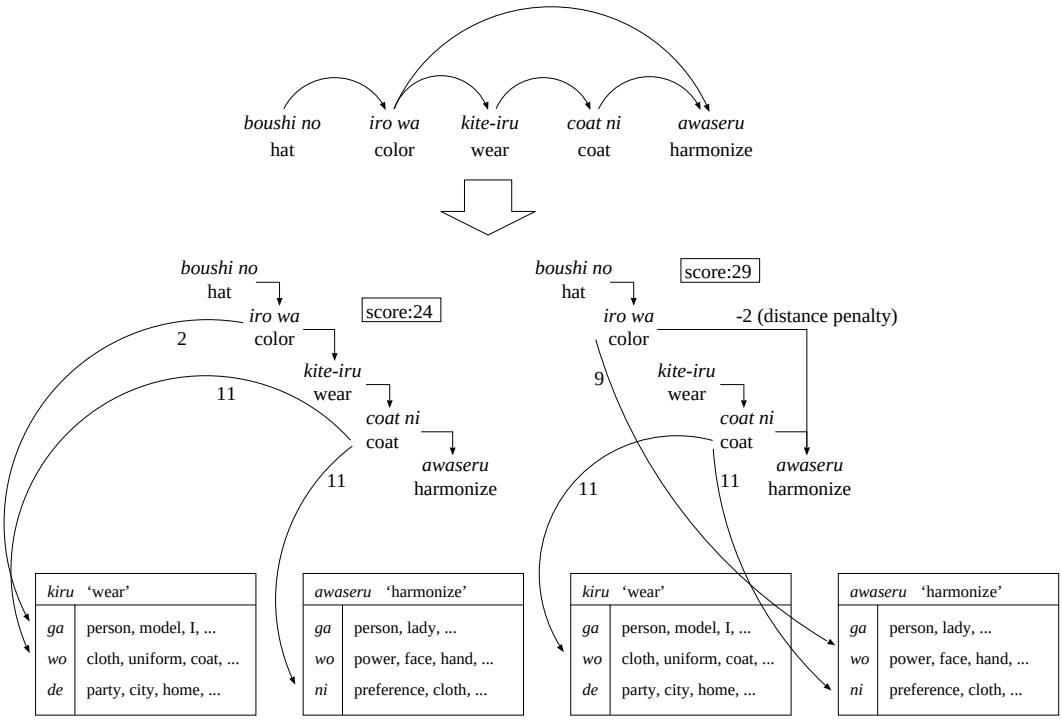

Figure 2: Parsing with case structure analysis.

sum of the matching scores in it.

Since the heuristic rule employed in KNP is actually very useful, we incorporate it, that is, penalty score is imposed to the modifier-head relation depending on the distance between a modifier and a head. If a modifier is not followed by a comma, the penalty score, $0, -2, -4, -6$, ... is imposed when a modifier modifies the first (nearest), second, third, fourth, ... verbs in a sentence respectively; if with a comma, the penalty score, $-2, 0, -2, -4$, ... is imposed.

For example, sentence 3 was analyzed by our method as shown in Figure 2. Since the similarity score between *iro* 'color' and the *wo*-slot of *awaseru* 'harmonize' is much larger than that between *iro* 'color' and the *ga*-slot of *kiru* 'wear', the correct structure of the sentence was detected (the right-hand parse of Figure 2). Note that, furthermore, both the case of *iro* in relation to *awaseru* 'harmonize', and the case of *coat* in relation to *kite-iru* 'wear' were detected correctly.

Structural ambiguities often cause a combinatorial explosion when a sentence is long. However, by detecting the scopes of coordinate structures beforehand, which often appear in long sentences, we can reasonably limit the possible structures of the sentence.

Table 3: The accuracy of case detection.

	correct case detection	incorrect case detection	Parsing error
topic-marker	82	13	5
clausal modifier	73	18	9

The average analysis speed of the experiments described in the next section was about 50 sentences/min. The time-out of one min. was only employed to 7 out of 4,272 test sentences.

4.3 Experiments and discussion

We used 4,272 sentences of Kyoto University corpus as a test set. We parsed them by our new method (Figure 3 shows several examples) and checked two points: case detection of ambiguous case components and syntactic analysis.

First, we randomly selected ambiguous case components: 100 topic-marked case components and 100 case components modified by clausal

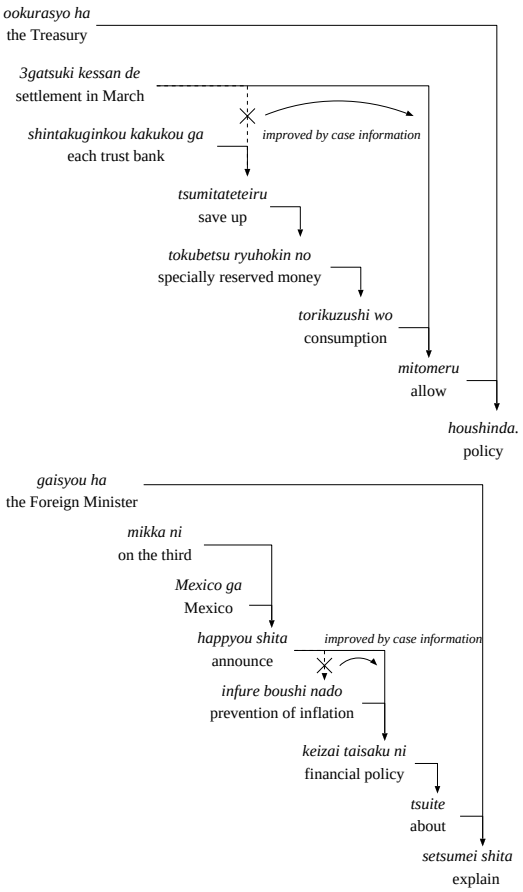

Figure 3: Examples of the analysis results.

modifiers, and checked whether their cases were correctly detected or not. As shown in Table 3, the accuracy of the analysis was fairly good: that for topic-markers was 82% and that for clausal modifiers was 73%.

Then, we compared the parse results of our method with those of the original KNP. As a result, 565 modifier-head relations differed; in 260 cases, our method was correct and the original KNP was incorrect (by considering the structures in the Kyoto University Corpus as a golden standard); in 224 cases, vice versa. That is, our method was superior to KNP by 36 cases, and increased the overall accuracy from 89.8% to 89.9%. Since the heuristic rule used in KNP is very strong, the improvement was not big. The improvement of the accuracy, though small, is valuable, because the accuracy around 90% seems close to the ceiling of this task.

5 Conclusion

We proposed an unsupervised construction method of a case frame dictionary. We obtained a large case frame dictionary, which consists of 23,497 verbs. Using this dictionary, we can detect ambiguous case components accurately. Also since our method employs unsupervised dictionary learning, it can be easily scaled up.

References

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of ANLP-97*.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentarou Ogura, and Yoshifumi Oyama Yoshihiko Hayashi, editors. 1997. *Japanese Lexicon*. Iwanami Publishing.

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

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

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

Christopher D. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of ACL-93*.

Takehito Utsuro, Takashi Miyata, and Yuji Matsumoto. 1998. General-to-specific model selection for subcategorization preference. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*.