Fertilization of Case Frame Dictionary for Robust Japanese Case Analysis

Daisuke Kawahara[†] and **Sadao Kurohashi**[†‡]

[†]Graduate School of Information Science and Technology, University of Tokyo
[‡]PRESTO, Japan Science and Technology Corporation (JST)
{kawahara,kuro}@kc.t.u-tokyo.ac.jp

Abstract

This paper proposes a method of fertilizing a Japanese case frame dictionary to handle complicated expressions: double nominative sentences, non-gapping relation of relative clauses, and case change. Our method is divided into two stages. In the first stage, we parse a large corpus and construct a Japanese case frame dictionary automatically from the parse results. In the second stage, we apply case analysis to the large corpus utilizing the constructed case frame dictionary, and upgrade the case frame dictionary by incorporating newly acquired information.

1 Introduction

To understand a text, it is necessary to find out relations between words in the text. What is required to do so is a case frame dictionary. It describes what kinds of cases each verb has and what kinds of nouns can fill a case slot. Since these relations have millions of combinations, it is difficult to construct a case frame dictionary by hand. We proposed a method to construct a Japanese case frame dictionary automatically by arranging large volumes of parse results by coupling a verb and its closest case component (Kawahara and Kurohashi, 2001). This case frame dictionary, however, could not handle complicated expressions: double nominative sentences, non-gapping relation of relative clauses, and case change.

This paper proposes a method of fertilizing the case frame dictionary to handle these complicated expressions. We take an iterative method which consists of two stages. This means gradual learning of what is understood by an analyzer in each stage. In the first stage, we parse a large raw corpus and construct a Japanese case frame dictionary automatically from the parse results. This is the method proposed by (Kawahara and Kurohashi, 2001). In the second stage, we apply case analysis to the large corpus utilizing the constructed case frame dictionary, and upgrade the case frame dictionary by incorporating newly acquired information.

We conducted a case analysis experiment with the upgraded case frame dictionary, and its evaluation showed effectiveness of the fertilization process.

2 Japanese Grammar

We introduce Japanese grammar briefly in this section.

Japanese is a head-final language. Word order does not play a case-marking role. Instead, postpositions function as case markers (CMs). The basic structure of a Japanese sentence is as follows:

(1)	*kare*	*ga*	*hon*	*wo*	*kaku*
	he	nom-CM	book	acc-CM	write

(he writes a book)

ga and *wo* are postpositions which mean nominative and accusative, respectively. *kare ga* and *hon wo* are case components, and *kaku* is a verb[1].

There are two phenomena that case markers are hidden.

A modifying clause is left to the modified noun in Japanese. In this paper, we call a noun modified by a clause **clausal modifiee**. A clausal modifiee is usually a case component for the verb of the modifying clause. There is, however, no case marker for their relation.

[1] In this paper, we call verbs, adjectives, and noun+copulas as verbs for convenience.

(2)

hon	*wo*	*kaita*	*hito*
book	acc-CM	write	person

(the person who wrote the book)

(3)

kare	*ga*	*kaita*	*hon*
he	nom-CM	write	book

(a book which he wrote)

In (2), *hito* 'person' has *ga* 'nominative' relation to *kaita* 'write'. In (3), *hon* 'book' has *wo* 'accusative' relation to *kaita* 'write'.

There are some non case-marking postpositions, such as *wa* and *mo*. They topicalize or emphasize noun phrases. We call them **topic markers (TMs)** and a phrase followed by one of them **TM phrase**.

(4)

kare	*wa*	*hon*	*wo*	*kaita*
he	TM	book	acc-CM	write

(he wrote a book)

(5)

kare	*ga*	*hon*	*mo*	*kaita*
he	nom-CM	book	TM	write

(he wrote a book also)

In (4), *wa* is interpreted as *ga* 'nominative'. In (5), *mo* is interpreted as *wo* 'accusative'.

3 Construction of the initial case frame dictionary

This section describes how to construct the initial case frame dictionary. This is the first stage of our two-stage approach, and is performed by the method proposed by (Kawahara and Kurohashi, 2001). In the rest of this section, we describe this approach in detail.

The biggest problem in automatic case frame construction is verb sense ambiguity. Verbs which have different meanings should have different case frames, but it is hard to disambiguate verb senses very precisely. To deal with this problem, we distinguish predicate-argument examples, which are collected from a large corpus, by coupling a verb and its closest case component. 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'.

This process makes separate case frames which have almost the same meaning or usage. For example, "*nimotsu wo tsumu*" 'load baggage' and "*busshi wo tsumu*" 'load supply' are separate case frames. To merge these similar case frames and increase coverage of the case frame, we cluster the case frames.

We employ the following procedure for the automatic case frame construction:

1. A large raw corpus is parsed by a Japanese parser, and reliable predicate-argument examples are extracted from the parse results. Nouns with a TM such as *wa* or *mo* and clausal modifiees are discarded, because their case markers cannot be understood by syntactic analysis.
2. The extracted examples are bundled according to the verb and its closest case component, making initial case frames.
3. The initial case frames are clustered using a similarity measure, resulting in the final case frames. The similarity is calculated by using NTT thesaurus.

We constructed a case frame dictionary from newspaper articles of 20 years (about 20,000,000 sentences).

4 Target expressions

The following expressions could not be handled with the initial case frame dictionary shown in section 3, because of lack of information in the case frame.

Non-gapping relation

This is the case in which the clausal modifiee is not a case component of the verb in the modifying clause, but is semantically associated with the clause.

(6)

kare ga	*syudoken wo*	*nigiru*	*kaigi*
he	initiative	have	meeting

(the meeting in which he has the initiative)

In this example, *kaigi* 'meeting' is not a case component of *nigiru* 'have', and there is no case relation between *kaigi* and *nigiru*. We call this relation **non-gapping relation**.

Double nominative sentence

This is the case in which the verb has two nominatives in sentences such as the following.

(7) *kuruma wa* *engine ga* *yoi*
car TM engine good
(the engine of the car is good)

In this example, *wa* plays a role of nominative, so *yoi* 'good' subcategorizes two nominatives: *kuruma* 'car' and *engine*. We call this outer nominative ***outer ga*** and this sentence **double nominative sentence**.

Case change

In Japanese, to express the same meaning, we can use different case markers. We call this phenomenon **case change**.

(8) *Tom ga* *Mary no* *shiji wo* *eta*
Tom Mary of support derive
(Tom derived his support from Mary)

In this example, *Mary* has *kara* 'from' relation to *eta* 'derive'. In this paper, we handle case change related to *no* 'of', such as (*no*, *kara*).

The following is an example that outer nominative is related to *no* case.

(9) *kuruma no* *engine ga* *yoi*
car engine good
(the engine of the car is good)

The outer nominative of (7) can be nominal modifier of the inner nominative like this example. This is case change of (*no*, *outer ga*).

There is a different case from the above that an NP with *no* modifying a case component does not have a case relation to the verb.

(10) *kare ga* *kaigi no* *syudoken wo* *nigiru*
he meeting initiative have
(he has the initiative in the meeting)

In this example, *kaigi* 'meeting' has a *no* relation to *syudoken* 'initiative', but does not have a case relation to *nigiru* 'have'. This example is a transformation of (6), and includes case change of (*no*, non-gapping).

5 Fertilization of case frame dictionary

We construct a fertilized case frame dictionary from the initial case frame dictionary shown in section 3, to handle the complicated expressions described in section 4.

We apply case analysis to a large corpus using the dictionary, collect information which could not be acquired by a mere parsing, and upgrade the case frame dictionary.

The procedure is as follows (figure 1):

1. The initial case frames are acquired by the method shown in section 3.
2. Case analysis utilizing the case frames acquired in phase 1 is applied to a large corpus, and examples of outer nominative are collected from case analysis results.
3. Case analysis utilizing the case frames acquired in phase 2 is applied to the large corpus, and examples of non-gapping relation are collected similarly.
4. Case similarities are judged to handle case change.

5.1 Case analysis based on the initial case frame dictionary

Case analysis of TM phrases and clausal modifiees is indebted to a case frame dictionary. This section describes an example of case analysis utilizing the initial case frame dictionary.

(11) *sono hon wa* *kare ga*
that book TM he
tosyokan de *yonda*
library in read
(he read that book in the library)

Case analysis of this example chooses the following case frame "*tosyokan de yonda*" 'read in the library' ("*" in the case frame means the closest CM.).

	CM	examples	input
read	nom	person, child, ···	he
	acc	book, paper, ···	book
	loc*	library, house, ···	library

kare 'he' and *tosyokan* 'library' correspond to nominative and locative, respectively, according to the surface cases. The case marker of TM phrase "*hon wa*" 'book (TM)' cannot be understood by the surface case, but it is interpreted as *wo* 'accusative' because of the matching between "*hon wa*" 'book (TM)' and the accusative case slot of the case frame (underlined in the case frame).

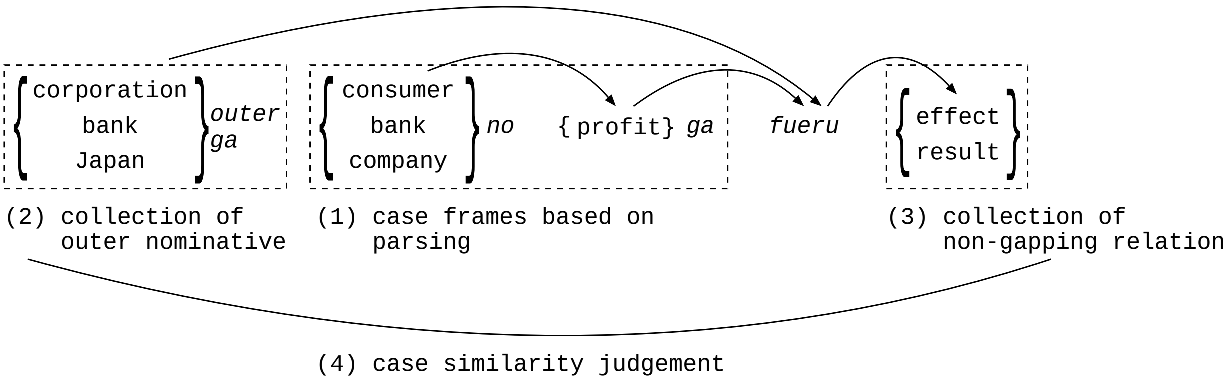
corporation
bank
Japan
outer ga
consumer
bank
company
no
{ profit} ga
fueru
effect
result
(2) collection of outer nominative
(1) case frames based on parsing
(3) collection of non-gapping relation
(4) case similarity judgement

Figure 1: Outline of our method

5.2 Collecting examples of outer nominative

In the initial case frame construction described in section 3, the TM phrase was discarded, because its case marker could not be understood by parsing. In the example (7), "*engine ga yoi*" 'the engine is good' is used to build the initial case frame, but the TM phrase "kuruma wa" 'the car' is not used.

Case analysis based on the initial case frame dictionary tells a case of a TM phrase. Correspondence to outer nominative cannot be understood by the case slot matching, but indirectly. If the TM cannot correspond to any case slots of the initial case frame, the TM can be regarded as outer nominative. For example, in the case of (7), since the case frame of "*engine ga yoi*" 'the engine is good' has only nominative which corresponds to "*engine*", the TM of "*kuruma wa*" cannot correspond to any case slots and is recognized as outer nominative. On the other hand, in the case of (11), the TM of *hon wa* is recognized as accusative, because *hon* 'book' is similar to the examples of the accusative slot. We can distinguish and collect outer nominative examples in this way.

We apply the following procedure to each sentence which has both a TM and *ga*. To reduce the influence of parsing errors, the collection process of these sentences is done under the condition that a TM phrase has no candidates of its modifying head without its verb.

1. We apply case analysis to a verb which is a head of a TM phrase. If the verb does not have the closest case component and cannot select a case frame, we quit processing this sentence and proceed to the next sentence. In this phase, the TM phrase is not made correspondence with a case of the selected case frame.
2. If the case frame does not have any cases which have no correspondence with the case components in the input, the TM cannot correspond to any case slots and is regarded as outer nominative. This TM phrase is added to outer nominative examples of the case frame.

The following is an example of this process.

(12)	*nagai*	*sumo*	*wa*	*ashi-koshi*	*ni*
	long	sumo	TM	legs and loins	
	futan	*ga*	*kakaru*		
	burden		impose		

(long sumo imposes a burden on legs and loins)

Case analysis of this example chooses the following case frame "*futan ga kakaru*" 'impose a burden'.

	CM	examples	input
impose	nom*	burden	burden
	dat	heart, legs, loins, · · ·	legs and loins

futan 'burden' and *ashi-koshi* 'legs and loins' correspond to nominative and dative of the case frame, respectively, and *sumo* corresponds to no case marker. Accordingly, the TM of "*sumo wa*" is recognized as outer nominative, and *sumo* is added to outer nominative examples of the case frame "*futan ga kakaru*".

This process made outer nominative of 15,302 case frames (of 597 verbs).

5.3 Collecting examples of non-gapping relation

Examples of non-gapping relation can be collected in a similar way to outer nominative. When a clausal modifiee has non-gapping relation, it should not be similar to any examples of any cases in the case frame, because the constructed case frames have examples of only cases except for non-gapping relation. From this point of view, we apply the following procedure to each example sentence which contains a modifying clause. To reduce the influence of parsing errors, the collection process of example sentences is done under the condition that a verb in a clause has no candidates of its modifying head without its clausal modifiee ("··· [modifying verb] N_1 *no* N_2" is not collected).

1. We apply case analysis to a verb which is contained by a modifying clause. If the verb does not have the closest case component and cannot select a case frame, we quit processing this sentence and proceed to the next sentence. In this phase, the clausal modifiee is not made correspondence with a case of the selected case frame.
2. If the similarity between the clausal modifiee and examples of any cases which have no correspondence with input case components does not exceed a threshold, this clausal modifiee is added to examples of non-gapping relation in the case frame. We set the threshold 0.3 empirically.

The following is an example of this process.

(13) *gyomu* *wo* *itonamu*
business carry on

menkyo *wo* *syutoku-shita*
license get

(ϕ got a license to carry on business)

Case analysis of this example chooses the following case frame "{*gyomu, business*} *wo itonamu*" 'carry on { work, business }'.

	CM	examples	input
carry on	nom	bank, company, ···	-
	acc*	work, business	business

Nominative of this case frame has no correspondence with a case component of the input, so the clausal modifiee, *menkyo* 'license', is checked whether it can correspond to nominative case examples. In this case, the similarity between *menkyo* 'license' and examples of nominative is not so high. Consequently, the relation of *menkyo* 'license' is recognized as non-gapping relation, and *menkyo* is added to examples of non-gapping relation in the case frame "{*gyomu, business*} *wo itonamu*".

(14) *ihouni* *denwa* *gyomu* *wo*
illegally telephone business

itonande-ita *utagai*
carry on suspect

(suspect that ϕ carried on telephone business illegally)

In this case, the above case frame is also selected. Since *utagai* 'suspect' is not similar to the nominative case examples, it is added to case examples of non-gapping relation in the case frame.

This process made non-gapping relation of 23,094 case frames (of 637 verbs).

Collecting examples of non-gapping relation for all the case frames

Non-gapping relation words which have wide distribution over verbs can be considered to have non-gapping relation for all the verbs or case frames. We add these words to examples of non-gapping relation of all the case frames. For example, 5 verbs have *menkyo* 'license' (example (13)) in their non-gapping relation, and 381 verbs have *utagai* 'suspect' (example (14)). We, consequently, judge *utagai* has non-gapping relation for all the case frames. We call such a word **global non-gapping word**.

We treated words which have non-gapping relation for more than 100 verbs as global non-gapping words. We acquired 128 global non-gapping words, and the following is the examples of them (in English).

possibility, necessity, result, course, case, thought, schedule, outlook, plan, chance, ···

5.4 Case similarity judgement

To deal with case change, we applied the following process to every case frame with outer nominative and non-gapping relation.

1. A similarity of every two cases is calculated. It is the average of similarities between all the combinations of case examples. But similarities of couples of basic cases are not handled, such as (*ga*, *wo*), (*ga*, *ni*), (*wo*, *ni*), and so on.
2. A couple whose similarity exceeds a threshold is judged to be similar, and is merged into one case. We set the threshold 0.8 empirically.

The following example is the case when this process is applied to "{*setsumei, syakumei*} *wo motomeru*" 'demand {explanation, excuse}'.

	CM	examples
demand	nom	committee, group, ⋯
	acc*	explanation, excuse
	dat	government, president, ⋯
	about	progress, condition, state, ⋯
	no	progress, reason, content, ⋯

In this case frame, the examples of *no* 'of'[2] are similar to those of *ni-tsuite* 'about', and the similarity between them is very high, 0.94, so these case examples are merged into a new case *no/ni-tsuite* 'of/about'.

By this process, 6,461 couples of similar cases are merged. An NP with *no* modifying a case component can be analyzed by this merging.

6 Case Analysis

To perform case analysis, we basically employ the algorithm proposed by (Kurohashi and Nagao, 1994). In this section, our case analysis method of the complicated expressions shown in section 4 is described.

6.1 Analysis of clausal modifiees

If an clausal modifiee is a function word such as *koto* '(that clause)' or *tame* 'due', or a time expression such as *3 ji* 'three o'clock' or *saikin* 'recently', it is analyzed as non-gapping relation.

The other clausal modifiee can correspond to *ga* 'nominative', *wo* 'accusative', *ni* 'dative', *outer ga* 'outer nominative', non-gapping relation, or *no* 'of'. We decide a corresponding case which maximizes the score[3] of the verb in the clause. If a clausal modifiee corresponds to *ga*, *wo*, *ni*, or *outer ga*, the relation is decided as it is. If it corresponds to non-gapping relation or *no*, the relation is decided as non-gapping relation. In the case of corresponding to *no*, the clausal modifiee has *no* relation to the closest case component of the verb.

A clausal modifiee can correspond to non-gapping relation or *no* under the condition that similarity between the clausal modifiee and case examples of non-gapping relation or *no* is the maximum value (which means two nouns locate in the same node in a thesaurus). This is because a noun which is a little similar to case examples of non-gapping relation may not have non-gapping relation.

6.2 Analysis of TM phrases

If a TM phrase is a time expression, it is analyzed as time case. The other TM phrase can correspond to *ga* 'nominative', *wo* 'accusative', or *outer ga* 'outer nominative'. We decide a corresponding case which maximizes the score of the verb modified by the TM phrase. When the verb has both a case component with *ga* and a TM phrase, the case component with *ga* corresponds to *ga* in the selected case frame, and its TM phrase corresponds to *wo* or *outer ga*. If the correspondence between the TM phrase and *outer ga* case components gets the best similarity, the input sentence is recognized as a double nominative sentence.

6.3 Analysis of case change

If the selected case frame of the input verb has merged cases which include *no* 'of', *no* case in the input sentence is interpreted as the counterpart of *no* between the merged cases. If not, the *no* case is considered not to have a case relation to the verb and has no corresponding case in the case frame.

[2] In *no* case in case frames, every noun which modifies the closest case component of the verb is collected.

[3] This score is the sum of each similarity between an input case component and examples of the corresponding case in the case frame.

Table 1: Case analysis accuracy

	clausal modifiee	TM
our method	301/358 84.0%	307/345 88.9%
baseline	287/358 80.1%	305/345 88.4%

Table 2: Non-gapping relation accuracy

	precision	recall	F
our method	82/116 70.7%	82/92 89.1%	78.8%
baseline	88/148 59.5%	88/92 95.7%	73.3%

7 Experiment

We made a case analysis experiment on Japanese relevance-tagged corpus (Kawahara et al., 2002). This corpus has correct tags of predicate-argument relations. We conducted case analysis on an open test set which consists of 500 sentences, and evaluated clausal modifiees and TM phrases in these sentences. To evaluate the real case analysis without influence of parsing errors, we input the correct structure of the corpus sentences to the analyzer.

The accuracy of clausal modifiees and TM phrases is shown in table 1, and the accuracy of non-gapping relation is shown in table 2. The baseline of these tables is that if a clausal modifiee belongs to a non-gapping noun dictionary in which nouns always having non-gapping relation as clausal modifiees are written, it is analyzed as non-gapping relation.

The accuracy of clausal modifiees increased by 4%. This shows effectiveness of our fertilization process. However, the accuracy of TM phrases did not increase. This is because the accuracy of TM phrases which were analyzed using added outer nominative examples was 4/6, and its frequency was too low. The accuracy of case change was 2/4.

8 Related work

There has been some related work analyzing clausal modifiees and TM phrases. Baldwin et al. analyzed clausal modifiees with heuristic rules or decision trees considering various linguistic features (Baldwin et al., 1999). Its accuracy was about 89%. Torisawa analyzed TM phrases using predicate-argument cooccurences and word classifications induced by the EM algorithm (Torisawa, 2001). Its accuracy was about 88% for *wa* and 84% for *mo*.

It is difficult to compare the accuracy because the range of target expressions is different. Unlike related work, it is promising to utilize our resultant case frame dictionary for subsequent analyzes such as ellipsis or discourse analysis.

9 Conclusion

This paper proposed a method of fertilizing the case frame dictionary to realize an analysis of the complicated expressions, such as double nominative sentences, non-gapping relation, and case change. We can analyze these expressions accurately using the fertilized case frame dictionary. So far, accuracy of subsequent analyzes such as ellipsis or discourse analysis has not been so high, because double nominative sentences and non-gapping relation cannot be analyzed accurately. It is promising to improve the accuracy of these analyzes utilizing the fertilized case frame dictionary.

References

Timothy Baldwin, Takenobu Tokunaga, and Hozumi Tanaka. 1999. The parameter-based analysis of Japanese relative clause constructions. In *IPSJ SIJ Notes 1999-NL-134*, pages 55–62.

Daisuke Kawahara and Sadao Kurohashi. 2001. Japanese case frame construction by coupling the verb and its closest case component. In *Proceedings of the Human Language Technology Conference*, pages 204–210.

Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. 2002. Construction of a Japanese relevance-tagged corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 2008–2013.

Sadao Kurohashi and Makoto Nagao. 1994. 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.

Kentaro Torisawa. 2001. An unsupervised method for canonicalization of Japanese postpositions. In *Proceedings of the 6th Natural Language Processing Pacific Rim Simposium*, pages 211–218.