Automatic Knowledge Acquisition for Case Alternation between the Passive and Active Voices in Japanese

Ryohei Sasano[1] **Daisuke Kawahara**[2] **Sadao Kurohashi**[2] **Manabu Okumura**[1]
[1] Precision and Intelligence Laboratory, Tokyo Institute of Technology
[2] Graduate School of Informatics, Kyoto University
{sasano,oku}@pi.titech.ac.jp, {dk,kuro}@i.kyoto-u.ac.jp

Abstract

We present a method for automatically acquiring knowledge for case alternation between the passive and active voices in Japanese. By leveraging several linguistic constraints on alternation patterns and lexical case frames obtained from a large Web corpus, our method aligns a case frame in the passive voice to a corresponding case frame in the active voice and finds an alignment between their cases. We then apply the acquired knowledge to a case alternation task and prove its usefulness.

1 Introduction

Predicate-argument structure analysis is one of the fundamental techniques for many natural language applications such as recognition of textual entailment, information retrieval, and machine translation. In Japanese, the relationship between a predicate and its argument is usually represented by using case particles[1] (Kawahara and Kurohashi, 2006; Taira et al., 2008; Yoshikawa et al., 2011). However, since case particles vary depending on the voices, we have to take case alternation into account to represent predicate-argument structure. There are thus two major types of representations: one uses surface cases, and the other uses normalized-cases for the base form of predicates. For example, while the Kyoto University Text Corpus (Kawahara et al., 2004), one of the major Japanese corpora that contains annotations of predicate-argument structures, adopts the former representation, the NAIST Text Corpora (Iida et al., 2007), another major Japanese corpus, adopts the latter representation.

Examples (1) and (2) describe the same event in the passive and active voices, respectively. When we use surface cases to represent the relationship between the predicate and its argument in Example (1), the case of "女 (woman)" is *ga*[2] and the case of "男 (man)" is *ni*.[2] On the other hand, when we use the normalized-cases for the base form, the case of "女 (woman)" is *wo*[2] and the case of "男 (man)" is *ga*, which are the same as the surface cases in the active voice as in Example (2).

(1) 女が (woman-*ga*) 男に (man-*ni*) 突き落とされた. (was pushed down)
(A woman was pushed down by a man.)

(2) 男が (man-*ga*) 女を (woman-*wo*) 突き落とした. (pushed down)
(A man pushed down a woman.)

Both representations have their own advantages. Surface case analysis is easier than normalized-case analysis, especially when we consider omitted arguments, which are also called zero anaphors (Nagao and Hasida, 1998). In Japanese, zero anaphora frequently occurs, and the omitted unnormalized-case of a zero anaphor is often the same as the surface case of its antecedent (Sasano and Kurohashi, 2011). Therefore, surface case analysis suits zero anaphora resolution. On the other hand, when

[1] Japanese is a head-final language. Word order does not mark syntactic relations. Instead, postpositional case particles function as case markers.

[2] *Ga*, *wo*, and *ni* are typical Japanese postpositional case particles. In most cases, they indicate nominative, accusative, and dative, respectively.

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1213–1223,
Seattle, Washington, USA, 18-21 October 2013.

we focus on the resulting predicate argument structures, the normalized-case structure is more useful. Specifically, since a normalized-case structure represents the same meaning in the same representation, normalized-case analysis is useful for recognizing textual entailment and information retrieval.

Therefore, we need a system that first analyzes surface cases and then alternates the surface cases with normalized-cases. In particular, we focus on the transformation of the passive voice into the active voice in this paper. Passive-to-active voice transformation in English can be performed systematically, which does not depend on lexical information in most cases. However, in Japanese, the method of transformation depends on lexical information. For example, while the case particle *ni* in Example (1) is alternated with *ga* in the active voice, the case particle *ni* in Example (3) is not alternated in the active voice as in Example (4) even though both their predicates are "突き落とされた (be pushed down)."

(3) 女が 海に 突き落とされた.
woman-*ga* sea-*ni* was pushed down
(A woman was pushed down into the sea.)

(4) 女を 海に 突き落とした.
woman-*wo* sea-*ni* pushed down
(ϕ pushed down a woman into the sea.)

The *ni* case in Example (1) indicates *agent*. On the other hand, the *ni* case in Example (3) indicates *direction*. To determine the difference is important for many NLP applications including machine translation. In fact, Google Translate (GT)[3] translates Examples (1) and (3) as "Woman was pushed down in the man" and "Woman was pushed down in the sea," respectively, which may be because GT cannot distinguish between the roles of *ni* in Examples (1) and (3).

(5) 賞が 男に 贈られた.
prize-*ga* man-*ni* was awarded
(A prize was awarded to a man.)

In example (5), although the *ni*-case argument "男 (man)" is the same as in Example (1), the case particle *ni* indicates *recipient* and is not alternated in the active voice. These examples show that case alternation between the passive and active voices in Japanese depends on not only predicates but also arguments, and we have to consider their combinations. Since it is impractical to manually describe the case alternation rules for all combinations of predicates and arguments, we have to acquire such knowledge automatically.

Thus, in this paper, we present a method for acquiring the knowledge for case alternation between the passive and active voices in Japanese. Our method leverages several linguistic constraints on alternation patterns and lexical case frames obtained from a large Web corpus, which are constructed for each meaning and voice of each predicate.

2 Related Work

Levin (1993) grouped English verbs into classes on the basis of their shared meaning components and syntactic behavior, defined in terms of diathesis alternations. Hence, diathesis alternations have been the topic of interest for a number of researchers in the field of automatic verb classification, which aims to induce possible verb frames from corpora (e.g., McCarthy 2000; Lapata and Brew 2004; Joanis et al. 2008; Schulte im Walde et al. 2008; Li and Brew 2008; Sun and Korhonen 2009; Theijssen et al. 2012). Baroni and Lenci (2010) used distributional slot similarity to distinguish between verbs undergoing the causative-inchoative alternations, and verbs that do not alternate.

There is some work on passive-to-active voice transformation in Japanese. Baldwin and Tanaka (2000) empirically identified the range and frequency of basic verb alternation, including active-passive alternation, in Japanese. They automatically extracted alternation types by using hand-crafted case frames but did not evaluate the quality. Kondo et al. (2001) dealt with case alternation between the passive and active voices as a subtask of paraphrasing a simple sentence. They manually introduced case alternation rules on the basis of verb types and case patterns and transformed passive sentences into active sentences.

Murata et al. (2006) developed a machine-learning-based method for Japanese case alternation. They extracted 3,576 case particles in passive sentences from the Kyoto University Text Corpus

[3] http://translate.google.com, accessed 2013-2-20.

Case particle	Grammatical function
ga	nominative
wo	accusative
ni	dative
de	locative, instrumental
kara	ablative
no	genitive

Table 1: Examples of Japanese postpositional case particles and their typical grammatical functions.

and tagged their cases in the active voice. Then, they trained SVM classifiers using the tagged corpus. Their features for training SVM were made by using several lexical resources such as IPAL (IPA, 1987), the Japanese thesaurus *Bunrui Goi Hyo* (NLRI, 1993), and the output of Kondo et al.'s method.

3 Lexicalized Case Frames

To acquire knowledge for case alternation, we exploit lexicalized case frames that are automatically constructed from 6.9 billion Web sentences by using Kawahara and Kurohashi (2002)'s method. In short, their method first parses the input sentences, and then constructs case frames by collecting reliable modifier-head relations from the resulting parses.

These case frames are constructed for each predicate like PropBank frames (Palmer et al., 2005), for each meaning of the predicate like FrameNet frames (Fillmore et al., 2003), and for each voice. However, neither pseudo-semantic role labels such as Arg1 in PropBank nor information about frames defined in FrameNet are included in these case frames. Each case frame describes surface cases that each predicate has and instances that can fill a case slot, which is fully lexicalized like the subcategorization lexicon VALEX (Korhonen et al., 2006).

We list some Japanese postpositional case particles with their typical grammatical functions in Table 1 and show examples of case frames in Table 2.[4] Ideally, one case frame is constructed for each meaning and voice of the target predicate. However, since Kawahara and Kurohashi's method is unsupervised, several case frames are actually constructed

[4] *Niyotte* in Table 2 is a Japanese functional phrase that indicates *agent* in this case. We treat *niyotte* as a case particle in this paper for the sake of simplicity.

Case Frame: "突き落とされる-4 (be pushed down-4)"
{ 女性 (woman):5, 僕 (I):2, 女 (woman):2, · · · }-*ga*
{ 海 (sea):229, 川 (bottom):115, 池 (pond):51, · · · }-*ni*
{ 継母(stepmother):2, ペガサス(Pegasus):2, · · · }-*niyotte*
· · ·

Case Frame: "突き落とされる-5 (be pushed down-5)"
{ 京子 (Kyoko):3, 監督 (manager):1, · · · }-*ga*
{ 誰か (someone):143, 何者か (somebody):85, · · · }-*ni*
{ 階段 (stair):20, 船 (ship):7, 崖 (cliff):7, · · · }-*kara*
· · ·

Case Frame: "突き落とす-2 (push down-2)"
{ 男 (man):14, 獅子 (lion):5, 虎 (tiger):3, · · · }-*ga*
{ 子(child):316, 子供(child):81, 人(person):51, · · · }-*wo*
{ 海 (sea):580, 谷 (ravine):576, 川 (river):352 · · · }-*ni*
· · ·

Case Frame: "突き落とす-4 (push down-4)"
{ 誰か (someone):14, ライオン (lion):5, · · · }-*ga*
{ 人 (person):257, 私 (I):214, 子 (child):137, · · · }-*wo*
{ 崖 (cliff):53, 階段 (stair):28, · · · }-*kara*
· · ·

Table 2: Examples of case frames for "突き落とされる (be pushed down)" and "突き落とす (push down)." Words in curly braces denote instances that can fill corresponding cases and the numbers following these words denote their frequency in the corpus.

for each meaning and voice. For example, 59 and eight case frames were respectively constructed for the predicate in the passive voice "突き落とされる (be pushed down)" and in the active voice "突き落とす (push down)" from 6.9 billion Web sentences. Table 2 shows the 4th and 5th case frames for "突き落とされる (be pushed down)" and the 2nd and 4th case frames for "突き落とす (push down)."

Table 3 shows an example of case frames for "殴る (hit)," which includes *no*-case. Here, the Japanese postpositional case particle "*no*" roughly corresponds to "of," that is, "*X no Y*" means "Y of X," and thus *no*-case is not an argument of the target predicate. While Kawahara and Kurohashi's method basically collects arguments of the target predicate, the phrase of *no*-case that modifies the direct object of the predicate is also collected as *no*-case. This is because, as we will show in the next section, this phrase can be represented as *ga*-case in the passive voice.

Case Frame: “殴る-2 (hit-2)”
{ 男 (man):51, 拳 (fist):30, 誰か (someone):23, · · · }-*ga*
{ 自分 (myself):360, 私 (I):223, · · · }-*no*
{ 頭 (head):5424, 顔 (face):3215, · · · }-*wo*
{ 拳 (fist):316, 平手 (palm):157, 拳骨 (fist):126, · · · }-*de*
· · ·

Table 3: An example of case frames for “殴る (hit).”

4 Passive-Active Transformation in Japanese

Morphologically speaking, the passive voice in Japanese is expressed by using the auxiliary verbs “れる (*reru*)” and “られる (*rareru*),” whose past forms are “れた (*reta*)” and “られた (*rareta*),” respectively. For example, the verb in the base form “突き落とす (*tsukiotosu*, push down)” is transformed into the past passive form “突き落とされた (*tsukiotosa-reta*, was pushed down).” Case alternations accompany passive-active transformation in Japanese. There are only two case alternations at most in passive-active transformation. One is the case represented as *ga* in the passive voice, and the other is the case represented as *ga* in the active voice.

Japanese passive sentences can be classified into three types in accordance with what is represented as *ga*-case in the passive voice: **direct passive**, **indirect passive**, and **possessor passive**.

In **direct passive** sentence, the object of the predicate in the active voice is represented as *ga*-case. Examples (1), (3), and (5) are all direct passive sentences. The case that is represented as *ga* in the active voice is usually represented as *ni*, *niyotte*, *kara*, or *de* in the passive sentence. In the first sentence of Examples (6) and (7),[5] *ga*-cases in the active voice are represented as *niyotte* and *kara*, respectively. On the other hand, *ga*-case in the passive sentence is alternated with *wo* or *ni* as shown with broken lines in the second sentence of Examples (6) and (7).

(6) P: 原因が 男によって 特定された.
cause-*ga* man-*niyotte* was identified
(The cause was identified by a man.)

A: 男が 原因を 特定した.
man-*ga* cause-*wo* identified
(A man identified the cause.)

[5]“P” denotes a passive sentence and “A” denotes the corresponding active sentence in these examples.

(7) P: 男が 女から 話しかけられた.
man-*ga* woman-*kara* was talked to
(A man was talked to by a woman.)

A: 女が 男に 話しかけた.
woman-*ga* man-*ni* talked to
(A woman talked to a man.)

Indirect passive is also called adversative passive, in which an indirectly influenced agent is represented with *ga*. For example, “私 (I),” the argument represented with *ga* in the first sentence of Example (8), does not appear in the active voice, i.e. the second sentence of Example (8). In the case of indirect passive, *ga*-case in the active sentence is always alternated with *ni*-case in the passive sentence as shown with solid lines in Examples (8).

(8) P: 私が 子供に 泣かれた.
I-*ga* child-*ni* was cried
(I’ve got a child crying.)

A: 子供が 泣いた. (A child cried.)
child-*ga* cried

Possessor passive is similar to indirect passive in that the argument represented with *ga*-case does not appear as an argument of the predicate in the active voice. Therefore, possessor passive is sometimes treated as a kind of indirect passive. However, in the case of possessor passive, the argument appears in the active sentence as a possessor of the direct object. For example, the *ga*-case argument “女 (woman)” in the passive sentence of Example (9) does not appear as an argument of the predicate “殴った (hit)” in the active sentence but appears in the phrase that modifies the direct object “頭 (head)” with the case particle *no*, which indicates that “女 (woman)” is the possessor of “頭 (head).”

(9) P: 女が 男に 頭を 殴られた.
woman-*ga* man-*ni* head-*wo* was hit
(A woman was hit on the head by a man.)

A: 男が 女の 頭を 殴った.
man-*ga* woman-*no* head-*wo* hit
(A man hit the head of a woman.)

In conclusion, the number of case alternation patterns accompanying passive-active transformation in Japanese is limited. *Ga*-case in the passive voice can

be alternated only with either *wo*, *ni*, or *no*, or does not appear in the active voice. *Ga*-case in the active voice can be represented only by *ni*, *niyotte*, *kara*, or *de* in the passive voice. Hence, it is sufficient to consider only their combinations.

5 Knowledge Acquisition for Case Alternation

5.1 Task Definition

Our objective is to acquire knowledge for case alternation between the passive and active voices in Japanese. We leverage lexical case frames obtained from a large Web corpus by using Kawahara and Kurohashi (2002)'s method and align cases of a case frame in the passive voice and cases of a case frame in the active voice. As described in Section 2, several case frames are constructed for each voice of each predicate. Our task consists of the following two subtasks:

1. Identify a corresponding case frame in the active voice.
2. Find an alignment between cases of case frames in the passive and active voice.

Figure 1 shows the overview of our task. If a case frame in the passive voice is input, we identify a corresponding case frame in the active voice, and find an alignment between cases by using the algorithm described in Section 5.3. In this example, an active case frame "突き落とす-4 (push down-4)" is identified as a corresponding case frame for the input passive case frame "突き落とされる-5 (be pushed down-5)" and *ga*, *ni*, and *kara*-cases in the passive case frame are aligned to *wo*, *ga*, and *kara*-cases in the active case frame, respectively.

5.2 Clues for Knowledge Acquisition

We exploit three clues for corresponding case frame identification and case alignment as follows:

1. Semantic similarity between the instances of the aligned cases: sim_{SEM}.
2. Case distribution similarity between the corresponding case frames: sim_{DIST}.
3. Preference of alternation patterns: f_{PP}.

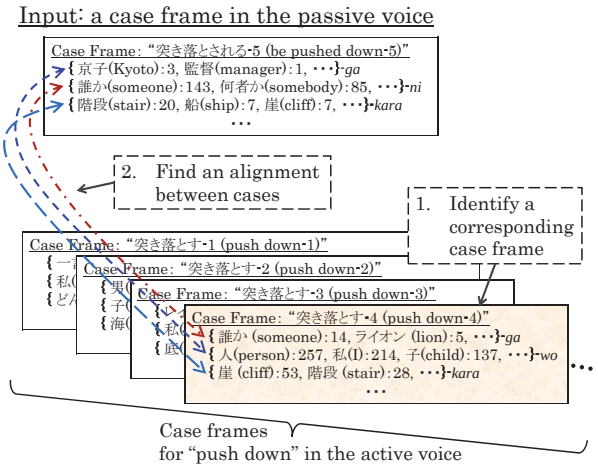

Figure 1: The overview of our task.

Semantic similarity The instances of the aligned cases should be similar. For example, the instances of the *ga*-case of the case frame "突き落とされる-5 (be pushed down-5)" and the *wo*-case of the case frame "突き落とす-4 (push down-4)," which are considered to be aligned and represent *patient*, are similar. Thus, we exploit semantic similarity sim_{SEM} between the instances of the corresponding cases.

We first define an asymmetric similarity measure between C_1 and C_2, each of which is a set of case slot instances, as follows:

$$\mathrm{sim}_a(C_1, C_2) = \frac{1}{|C_1|} \sum_{i_1 \in C_1} \max_{i_2 \in C_2} (\mathrm{sim}(i_1, i_2)),$$

where $\mathrm{sim}(i_1, i_2)$ is the similarity between instances. In this study, we apply a distributional similarity measure (Lin, 1998), which was computed from the Web corpus used to construct the case frames. We next define a symmetric similarity measure between C_1 and C_2 as an average of $\mathrm{sim}_a(C_1, C_2)$ and $\mathrm{sim}_a(C_2, C_1)$.

$$\mathrm{sim}_s(C_1, C_2) = \frac{1}{2}(\mathrm{sim}_a(C_1, C_2) + \mathrm{sim}_a(C_2, C_1)).$$

Then we define semantic similarity of a case alignment A between case frames CF_1 and CF_2.

$$\mathrm{sim}_{SEM}(A) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{sim}_s(C_{1,i}, C_{2,a(i)}),$$

where N denotes the number of case slots of CF_1, $C_{1,i}$ denotes a set of instances of the i-th case slot of CF_1, and $C_{2,a(i)}$ denotes the set of the aligned case instances of CF_2. A denotes the alignment $\{c_{1,1} \rightarrow c_{2,a(1)}, c_{1,2} \rightarrow c_{2,a(2)}, \ldots, c_{1,N} \rightarrow c_{2,a(N)}\}$ where $c_{n,i}$ denotes the case name that corresponds to $C_{n,i}$.

Case distribution similarity Although arguments are often omitted in Japanese, arguments that are usually mentioned explicitly in the passive voice will be also explicitly mentioned in the active voice. Hence, the frequency distribution of cases can be a clue for case alignment. In this study, we exploit the following cosine similarity of frequency distribution as case distribution similarity:

$$\mathrm{sim}_{DIST}(A) = \cos((|C_{1,1}|, \ldots, |C_{1,N}|), (|C_{2,a(1)}|, \ldots, |C_{2,a(N)}|)).$$

As an example, consider the alignment between a passive case[6] "選ばれる-1 (be selected-1)" and the corresponding active case frame "選ぶ-13 (select-13)" in Table 4. The alignment $A_1 = \{ga \rightarrow wo, ni \rightarrow ni, \mathrm{NIL} \rightarrow ga\}$ is considered to be correct. However, if we consider only the semantic similarity, an alignment $A_2 = \{ga \rightarrow ni, ni \rightarrow ga, wo \rightarrow wo\}$ is selected, because the alignment A_2 has the highest semantic similarity. On the other hand, the case distribution similarity

$$\mathrm{sim}_{DIST}(A_1) = \cos((17722, 122273, 0), (33338, 800, 382)) \approx 0.167$$

is much larger than

$$\mathrm{sim}_{DIST}(A_2) = \cos((17722, 122273, 96), (800, 382, 33338)) \approx 0.016.$$

Thus, the alignment A_1 would be selected by considering the case distribution similarity.

Preference of alternation patterns Some alternation patterns often appear, and others do not. For example, as Murata et al. (2006) reported, whereas 96.47% of *ga*-case is alternated with *wo*-case in passive-active transformation in Japanese, only 27.38% of *ni*-case is alternated with *ga*-case. Therefore, when we can use development data, we exploit a weighting factor $f_{PP}(A)$ that is determined on the development data and takes into account the preference of alternation patterns. We define $f_{PP}(A)$ as follows:

$$f_{PP}(A) = w(ga \rightarrow c_{ga\_to}) \times w(c_{to\_ga} \rightarrow ga), \quad \text{(i)}$$

where c_{ga_to} is the case in the active voice to which *ga*-case in the passive voice is aligned, c_{to_ga} is the case in the passive voice which is aligned to *ga*-case in the active voice, and $w(c_1 \rightarrow c_2)$ denotes the weight of the case alternation "$c_1 \rightarrow c_2$."

Case Frame: "選ばれる-1 (be selected-1)"
{ 選手 (player):1119, 作品 (work):983, $\cdots$ }-*ga*:17722
{ 代表 (representative):18295, $\cdots$ }-*ni*:122273
{ 作品 (work):5, 市長 (mayor):3, $\cdots$ }-*wo*:96
$\cdots$

Case Frame: "選ぶ-13 (select-13)"
{ 私 (I):14, 先生 (teacher):18, $\cdots$ }-*ga*:382
{ 優秀賞 (award):42, シングル (single):17, $\cdots$ }-*ni*:800
{ 曲 (tune):16666, 作品 (work):9967, $\cdots$ }-*wo*:33338
$\cdots$

Table 4: Case frames "選ばれる-1 (be selected-1)" and "選ぶ-13 (select-13)." The numbers following case names denote the total numbers of case slot instances.

5.3 Algorithm

Algorithm 1 presents our algorithm for identifying a corresponding case frame and finding an alignment between cases in pseudo-code. Our algorithm first makes all possible combinations of a case frame in the active voice (cf_{active}), a case in the active voice to which *ga*-case in the passive voice is aligned (c_{ga_to}), and a case in the passive voice which is aligned to *ga*-case in the active voice (c_{to_ga}) on the basis of the linguistic constraints, and then evaluates the score for the combinations $\{cf_{active}, c_{ga\_to}, c_{to\_ga}\}$ by the following equation:

$$score = \mathrm{sim}_{SEM}(A) \times \mathrm{sim}_{DIST}(A)^{\alpha} \times f_{PP}(A), \quad \text{(ii)}$$

where α is a parameter that controls the impact of the case distribution similarity.[7] When we can use

[6] This case frame should not have *wo*-case. However, since we constructed case frames automatically, some case frames have improper cases.

[7] Since $f_{PP}(A)$ is defined with a set of weights of case alternation patterns, $f_{PP}(A)$ contains these weights implicitly, and thus there is only a single explicit weight in equation (ii).

Algorithm 1: Identifying a corresponding case frame and finding an alignment between cases.

Input: a case frame in the passive voice: $cf_{passive}$, and a set of case frames in the active voice: CFS_{active}
Output: a case frame and an alignment between cases: A

1: $max\_score = 0$, $A = ()$
2: **for each** $cf_{active} \in CFS_{active}$
3: **for each** $c_{ga\_to} \in$ {*wo*, *ni*, *no*, NIL}
4: **for each** $c_{to\_ga} \in$ {*ni*, *niyotte*, *kara*, *de*, NIL}
5: **if** (!occur(c_{ga_to}, c_{to_ga})) **then continue**
6: $A' = (cf_{active}, c_{ga\_to}, c_{to\_ga})$
7: $score = \text{sim}_{SEM}(A') \times \text{sim}_{DIST}(A')^{\alpha} \times f_{PP}(A')$
8: **if** ($score > max\_score$) **then**
9: $(max\_score, A) = (score, A')$
10: **end for**
11: **end for**
12: **end for**

development data, we tune α on the development data; otherwise we set $\alpha = 1$. Since some combinations of c_{ga_to} and c_{to_ga} never occur, our algorithm filters them out in line 5 of the algorithm. After checking all combinations, the combination with the highest score is output.

6 Evaluation of the Acquired Knowledge

We applied our algorithm to the case frames that are automatically constructed from a corpus consisting of about 6.9 billion Japanese sentences from the Web. Of course, these case frames contain improper ones, that is, several frames mix several meanings or usages of the predicates. Thus, it is difficult to evaluate the acquired knowledge itself. Instead, we evaluate the usefulness of the acquired knowledge on a case alternation task between the passive and active voices.

6.1 Setting and Algorithm for Case Alternation

We basically used the same data as Murata et al. (2006). As mentioned in Section 2, they extracted 3,576 case particles in passive sentences from the Kyoto University Text Corpus, and tagged their cases in the active voice. Since they treated possessor passive as a kind of indirect passive, they did not adopt the case alternation between *ga* and *no*. In addition, their data included some annotation errors. We thus modified 21 annotations,[8] five of which were changed to the case alternation between *ga* and *no*. Note that there were some cases where multiple possible case particles were tagged to one instance. We adopted evaluation metrics called "Eval. B" by Murata et al., that is, we judged the output to be correct when the output was included in possible answers. We performed experiments on the following three types of data settings.

1. Experiments without either development or training data.

2. Experiments with development data.

3. Experiments with training data.

Experiments without either development or training data In the first setting, we aligned the input passive case frame to one of the active case frames of the same predicate only by using sim_{SEM} and sim_{DIST} with the parameter $\alpha = 1$. Therefore, this setting is fully unsupervised. In this setting, the input surface cases are alternated as follows:

1. If a passive sentence is input, perform syntactic and surface case structure analysis by using Kawahara and Kurohashi (2006)'s model.[9] Their model identified a proper case frame for each predicate, and assigned arguments in the input sentence to case slots of the case frame.

2. By using the acquired knowledge for case alternation, alternate input surface cases with cases in the active voice.

We call this model Model 1. For example, if Example (10) is input, the *ga*-case argument is assigned to the *ga*-case of the case frame "突き落とされる-5 (be pushed down-5)." Since this case is aligned to the *wo*-case of the case frame "突き落とす-4 (push down-4)" as shown in Figure 1, this *ga*-case is alternated with *wo*-case.

(10) 女が　突き落とされた.
woman-*ga* was pushed down
(A woman was pushed down.)

[8] The modified version of the data is publicly available at http://alaginrc.nict.go.jp/case/src/kaku1.1.tar.gz.

[9] KNP: http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

Algorithm 2: Pseudo-code of the hill-climbing algorithm for tuning the parameter vector $\mathbf{x}$.

$\mathbf{x} = (1.0, 1.0, \ldots, 1.0)$
$acc = f_{accuracy}(\mathbf{x}), pre\_acc = 0$
while $acc > pre\_acc$
  $pre\_acc = acc$
  for $i \in \{0, \ldots, |\mathbf{x}| - 1\}$
    $acc_+ = f_{accuracy}(x_0, \ldots, x_i + 0.1, \ldots, x_{|\mathbf{x}|-1})$
    $acc_- = f_{accuracy}(x_0, \ldots, x_i - 0.1, \ldots, x_{|\mathbf{x}|-1})$
    if $acc_+ > acc$ **and** $acc_+ > acc_-$ **then** $x_i = x_i + 0.1$
    else if $acc_- > acc$ **then** $x_i = x_i - 0.1$
    $acc = f_{accuracy}(\mathbf{x})$
  end for
end while

	Model			Accuracy
	sim_{SEM}	sim_{DIST}	Parameter tuning	
Model 1_S	✓			0.902 (3,224/3,576)
Model 1_D		✓		0.857 (3,063/3,576)
Model 1	✓	✓		0.906 (3,239/3,576)
Model 2_S	✓		✓	0.928 (3,320/3,576)
Model 2_D		✓	✓	0.927 (3,314/3,576)
Model 2	✓	✓	✓	0.938 (3,353/3,576)
Baseline				0.883 (3,159/3,576)

Table 5: Experimental results of case alternation without training data.

Experiments with development data In the second setting, we aligned the input passive case frame to one of the active case frames of the same predicate by using sim_{SEM}, sim_{DIST}, and f_{PP} with α tuned on the development data. In advance, we divided the tagged data into two parts just as Murata et al. (2006) did, both of which contained 1,788 case particles, and performed 2-fold cross-validation. We used one part for development and the other for testing, and vice versa.

We tuned $w(ga \rightarrow c_{ga\_to})$, $w(c_{to\_ga} \rightarrow ga)$ in Equation (i), and α in Equation (ii) by a simple hill-climbing strategy. Since the candidate cases for c_{ga_to} are *ni*, *niyotte*, *kara*, *de*, and NIL, and the candidate cases for c_{to_ga} are *wo*, *ni*, *no*, and NIL, we defined parameter vector $\mathbf{x}$ as follows:

$$\mathbf{x} = (w(ga \rightarrow ni), w(ga \rightarrow niyotte), w(ga \rightarrow kara), w(ga \rightarrow de), w(ga \rightarrow \text{NIL}), w(wo \rightarrow ga), w(ni \rightarrow ga), w(wo \rightarrow no), w(\text{NIL} \rightarrow ga), \alpha).$$

Algorithm 2 shows the hill-climbing algorithm for tuning the parameter vector $\mathbf{x}$, where $f_{accuracy}(\mathbf{x})$ is a function that returns the case alternation accuracy on the development data with parameter $\mathbf{x}$. This algorithm varies one parameter at a time with a step-size of 0.1 until there is no accuracy improvement in the development data. After acquiring knowledge for case alternation with the tuned parameter, we applied the same method for case alternation as the first setting. We call this model Model 2.

Experiments with training data In the third setting, we also performed 2-fold cross validation, that is, we used one part of the divided tagged corpus for training and the other for testing, and vice versa. Although we basically applied Murata et al. (2006)'s method, which is based on SVMs, we added the output of Model 2 as a new feature.

Specifically, we first tuned the parameter vector $\mathbf{x}$ on the training data and acquired the knowledge for case alternation with the tuned parameter. By using the acquired knowledge, we alternated the input cases in both the training and test data and obtained the resulting case of Model 2. Note that, we did not use any annotations for the test data in this process. We then trained the SVMs on the training data and applied them to the test data using the resulting case as a new feature. We call this model Model 3.

6.2 Results and Discussion

Table 5 shows the results of the experiments without training data. Baseline is a system that outputs the most frequently alternated cases in the development data, which was also used by Murata et al. (2006). The baseline score was higher than that reported by Murata et al. because we modified 21 annotations. We also performed experiments without using case distribution similarity or semantic similarity. We call these models in the first setting Model 1_S and Model 1_D, and these models in the second setting Model 2_S and Model 2_D, respectively.

Although Models 1_S, 1_D, and 1 were fully unsupervised models, Models 1_S and 1 significantly[10] outperformed the baseline model (p-values of McNemar (1947)'s test were smaller than 0.00001). On the other hand, the difference between Models 1_S

[10] In this paper, we call a difference significant if the p-value of McNemar (1947)'s test is less than 0.01.

Model	Accuracy
(Murata et al., 2006)	0.944 (3,376/3,576)
Model 3	0.956 (3,417/3,576)

Table 6: Comparison between Murata et al. (2006)'s method and our method with training data.

and 1 is not statistically significant, and thus the effect of the case distribution similarity was not confirmed by these experiments.

Models 2_S, 2_D, and 2 were models with parameter tuning. Parameter tuning significantly improved the performance. In addition, the difference between Models 2_S and 2 and the difference between Models 2_D and 2 were both significant (p-values of McNemar's test were 0.00032 and 0.00039, respectively), and thus we confirmed the usefulness of the two similarity measures. The parameter α that controls the impact of the case distribution similarity was tuned to 0.3, which means semantic similarity between the instances of the aligned cases is more important than case distribution similarity for this task.

Table 6 compares Murata et al.'s method and our method with training data. We used Murata et al.'s method without feature selection because it achieved the highest performance on this setting. Their method's score was higher than that they reported, again due to the corpus modification. The difference between their method and our method was significant (p-value of McNemar's test was 0.00011), and we confirmed the usefulness of the acquired knowledge for case alternation.

Table 7 shows an example of case alternation between the passive and active voices. When the passive sentence was input, the argument "松樹さんが(Mr. Matsuki-*ga*)" was first assigned to *ga*-case of the case frame "殴られる-2 (be hit-2)." Since this case was aligned to *no*-case of the case frame "殴る-2 (hit-2)," the input *ga*-case was alternated with *no*-case. On the other hand, the cases of the other arguments "バットで (bat-*de*)" and "頭を (head-*wo*)" were output as they were in the passive sentence.

We now list three error causes observed in our experiments of the case alternation task:

1) The passive voice in Japanese is expressed by using the auxiliary verbs "れる (*reru*)" and "られる (*rareru*)." However, these auxiliary verbs can represent several other meanings, such as honorific and possibility. Since Kawahara and Kurohashi (2002)'s method does not distinguish between these meanings, our case frames sometimes contain improper cases such as *wo*-case in case frame "選ばれる-1 (be selected-1)" in Table 4.

Input Text:
· · · 松樹さんが　金属バットで　頭を　殴られ、· · ·
Mr. Matsuki-*ga* metal bat-*de* head-*wo* was hit
(. . . Mr. Matsuki was hit on the head with a metal bat . . .)

Identified passive case frame:
Case Frame: "殴られる-2 (be hit-2)"
{ 何者か (someone):2, 部員 (member):1, · · · }-*niyotte*
{ 女性 (woman):5, 女児 (girl)):4, · · · }-***ga***
{ 頭 (head):3944, 顔 (face):1186, · · · }-*wo*
{ 鈍器 (blunt weapon):84, バット (bat):45, · · · }-*de*
· · ·

Corresponding active case frame and case alignment:
Case alignment: {*niyotte*→*ga*, ***ga*** **→*no***, *wo*→*wo*, *de*→*de*}
Case Frame: "殴る-2 (hit-2)"
{ 男 (man):51, 拳 (fist):30, 誰か (someone):23, · · · }-*ga*
{ 自分 (myself):360, 私 (I):223, · · · }-***no***
{ 頭 (head):5424, 顔 (face):3215, · · · }-*wo*
{ 拳 (fist):316, 平手 (palm):157, 拳 (fist):43, · · · }-*de*
· · ·

Table 7: An example of case alternation. The input *ga*-case was alternated with *no*-case.

2) In some passive sentences, there are two surface *ni*-cases as in Example (11). However, our method does not assume such sentences, and thus cannot deal with them properly.

(11) 男に (man-*ni*) オフィスに (office-*ni*) 派遣された. (was sent)
(ϕ was sent to the office by a man.)

3) *Agent* of a predicate can be represented by using several types of case particles in the passive voice. For example, "会社 (company)" in Example (12) is the *agent* of "雇用した (employed)," which can be represented by either of *ni*, *niyotte*, and *kara* in the passive voice. Since Kawahara and Kurohashi (2002)'s method can not recognize the exchangeablity of case particles, some case frames contain several cases of the same semantic role. However, since our method enforces a one-to-one alignments, only one of these cases is properly aligned to the corresponding case in the active voice.

(12) 会社 が　　男を　　雇用した.
company-*ga* man-*wo* employed
(The company employed a man.)

6.3 Application to Alternation between the Causative and Active Voices

To confirm the applicability of our framework to other types of alternation than the active-passive alternation, we applied our framework to case alternation between the causative and active voices. The causative voice in Japanese is a grammatical voice and is expressed by using the auxiliary verbs "せる (*seru*)" and "させる (*saseru*)." We basically used the same algorithm as Algorithm 1 for acquiring the knowledge for case alternation, but used different constraints on case alternation patterns because possible case alternation patterns are different from those of active-passive alternation. Specifically, we replaced the third and fourth lines of Algorithm 1 with "for each $c_{to\_ga} \in \{\text{NIL}, ni\}$" and "for each $c_{ga\_to} \in \{wo, ni\}$," respectively, based on linguistic analysis of active-causative alternation in Japanese.

We used a part of the data created by Murata and Isahara (2003) to evaluate the usefulness of the acquired knowledge. Their data consists of 4,671 case particles in passive or causative sentences from the Kyoto University Text Corpus with their cases in the active voice. We first extracted 524 case particles that were extracted from causative sentences. Since the annotation quality was not very high, we manually checked all tags and modified inappropriate ones. We then performed 2-fold cross validation experiments. Table 8 shows experimental results. Baseline is a system that outputs the most frequently alternated cases in the training data. The difference between Murata et al. (2006)'s model[11] and our method was significant (p-value of McNemar's test was 0.0019), and we confirmed the applicability of our framework to active-causative alternation.

Model	Accuracy
Baseline	0.781 (409/524)
Murata et al. (2006)'s model	0.836 (438/524)
Our method with training data	0.872 (457/524)

Table 8: Experimental results of case alternation between the causative and active voices.

7 Conclusions and Future Directions

We have presented a method for automatically acquiring knowledge for case alternation between the passive and active voices in Japanese. Our method aligned an input case frame in the passive voice to a corresponding case frame in the active voice and found an alignment between their cases. We then applied the acquired knowledge to a case alternation task and proved its usefulness.

The knowledge we have to manually construct is only the knowledge of linguistic constraints on case alternation patterns. The other types of knowledge are automatically acquired from a large raw corpus. Thus, although this paper focused on the active-passive alternation in Japanese, our framework is applicable to the other types of case alternation and to other languages, especially similar languages such as Korean. We plan to apply our framework to other types of case alternation such as case alternation between intransitive and transitive verbs.

[11] In this experiment, we used the same features as those used by Murata and Isahara (2003).

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 23800025 and 25730131.

References

Timothy Baldwin and Hozumi Tanaka. 2000. Verb alternations and Japanese – how, what and where? In *Proc. of PACLIC 14*, pages 3–14.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistic*, 36(4):673–721.

Charles J. Fillmore, Christopher R. Johnson, and Miriam. R. L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proc. of ACL'07 Workshop: Linguistic Annotation Workshop*, pages 132–139.

IPA. 1987. Japanese verbs : A guide to the IPA lexicon of basic Japanese verbs.

Eric Joanis, Suzanne Stevenson, and David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3):337–367.

Daisuke Kawahara and Sadao Kurohashi. 2002. Fertilization of case frame dictionary for robust Japanese case analysis. In *Proc. of COLING'02*, pages 425–431.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proc. of HLT-NAACL'06*, pages 176–183.

Daisuke Kawahara, Ryohei Sasano, and Sadao Kurohashi. 2004. Toward text understanding: Integrating relevance-tagged corpora and automatically constructed case frames. In *Proc. of LREC'04*, pages 1833–1836.

Keiko Kondo, Satoshi Sato, and Manabu Okumura. 2001. Paraphrasing by case alternation (in Japanese). *Journal of Information Processing Society of Japan*, 42(3):465–477.

Anna Korhonen, Yuval Krymolowski, and Ted Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. In *Proc.of LREC'06*, pages 3000–3006.

Mirella Lapata and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1):45–73.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Jianguo Li and Chris Brew. 2008. Which are the best features for automatic verb classification. In *Proc. of ACL-HLT'08*, pages 434–442.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. of ACL-COLING'98*, pages 768–774.

Diana McCarthy. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proc. of NAACL'00*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157.

Masaki Murata and Hitoshi Isahara. 2003. Conversion of Japanese passive/causative sentences into active sentences using machine learning. In *Proc. of CICLing'03*, pages 115–125.

Masaki Murata, Toshiyuki Kanamaru, Tamotsu Shirado, and Hitoshi Isahara. 2006. Machine-learning-based transformation of passive Japanese sentences into active by separating training data into each input particle. In *Proc. of COLING-ACL'06*, pages 587–594.

Katashi Nagao and Koiti Hasida. 1998. Automatic text summarization based on the global document annotation. In *Proc. of ACL'98*, pages 917–921.

NLRI. 1993. *Bunrui Goi Hyo (in Japanese)*. Shuuei Publishing.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1):71–105.

Ryohei Sasano and Sadao Kurohashi. 2011. A discriminative approach to japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proc. of IJCNLP'11*, pages 758–766.

Sabine Schulte im Walde, Christian Hying, Christian Scheible, and Helmut Schmid. 2008. Combining EM training and the MDL principle for an automatic verb classification incorporating selectional preferences. In *Proc. of ACL-HLT'08*, pages 496–504.

Lin Sun and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proc. of EMNLP'09*, pages 638–647.

Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. 2008. A Japanese predicate argument structure analysis using decision lists. In *Proc. of EMNLP'08*, pages 523–532.

Daphne Theijssen, Lou Boves, Hans van Halteren, and Nelleke Oostdijk. 2012. Evaluating automatic annotation: automatically detecting and enriching instances of the dative alternation. *Language Resources and Evaluation*, 46(4):565–600.

Katsumasa Yoshikawa, Masayuki Asahara, and Yuji Matsumoto. 2011. Jointly extracting Japanese predicate-argument relation with markov logic. In *Proc. of IJCNLP'11*, pages 1125–1133.