Acquiring Strongly-related Events using Predicate-argument Co-occurring Statistics and Case Frames

Tomohide Shibata and **Sadao Kurohashi**
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
{shibata,kuro}@i.kyoto-u.ac.jp

Abstract

This paper proposes a method for automatically acquiring strongly-related events from a large corpus using predicate-argument co-occurring statistics and case frames. The strongly-related events are acquired in the form of strongly-related two predicates with their relevant arguments. First, strongly-related events are acquired from predicate-argument co-occurring statistics. Then, the remaining argument alignment is performed by using case frames. We conducted experiments using a Web corpus consisting of 1.6G sentences. The accuracy for the extracted event pairs was 96%, and the accuracy of the argument alignment was 79%. The number of acquired event pairs was about 20 thousands.

1 Introduction

Natural language understanding requires a wide variety of knowledge. One is the relation between predicate and argument. This relation has been automatically acquired in the form of case frames from a large corpus, and is utilized for parsing (Kawahara and Kurohashi, 2006). Another is the relation between events. The relation between events includes temporal relation, causality, and so on, and is useful for coreference resolution (Bean and Riloff, 2004) and anaphora resolution (Gerber and Chai, 2010).

This paper extracts two strongly-related events. Since the meaning of a predicate itself is often ambiguous, an event is treated as predicate-argument structure, namely the predicate with their relevant arguments. An example of two strongly-related events is shown below[1]:

PA_1		PA_2
P_1: pick up		P_2: bring
nom A_1:{man, person}	$\Rightarrow$	nom A_1:{man, person}
acc A_2:{purse}		acc A_2:{purse}
		dat A_3:{police}

In the above example, while the argument A_1 and A_2 appear both in PA_1 and PA_2, the argument A_3 appears only in PA_2. The argument A_3 works for specifying the meaning of the predicate P_2. The method that automatically extracts sets of events from unlabeled corpora (Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009) relies on the coreference relation of arguments, and thus cannot extract an argument such as A_3.

In languages where an argument is often omitted, such as Japanese, sentences illustrating the above two events usually occur in the following form (for simplicity, the sentences are explained in English):

(1) a. A man picked up a purse and brought (ϕ) to the police.

b. (ϕ) picked up a purse and brought (ϕ) to the police.

In the sentence (1-a), the argument A_1 and A_2 are omitted in PA_2. Moreover, as an agent is specifically omitted, in the sentence (1-b), the argument A_1 in PA_1 is also omitted. The coreference-based method (Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009) is hard to be applied to such a language since an argument rarely appears in both PA_1 and PA_2.

Our proposed method extracts strongly-related events in a two-phrase construct. First, since the arguments, such as A_2 and A_3, which specify the meaning of the predicate occur in at least one predicate-argument structure, the co-occurrence measure between "pick up purse" and "bring to police" can be calculated from their occurrences. Thus, we can regard "pick up purse" and "bring to police", whose mutual information is high, as strongly-related events.

[1] nom, acc, dat denotes nominative, accusative, dative, respectively.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1028–1036,
Chiang Mai, Thailand, November 8 – 13, 2011.

Next, we identify the remaining arguments by using case frames (Kawahara and Kurohashi, 2006). Case frames describe what kinds of arguments each predicate takes and what kinds of nouns can fill a case slot. With the similarity of noun distribution between an argument in a case frame assigned to PA_1 and an argument in a case frame assigned to PA_2, the remaining arguments can be aligned. In the above example, acc A_2 ("purse") in PA_1 corresponds to acc A_2 in PA_2, and nom A_1 ("man", "person") in PA_1 corresponds to nom A_1 in PA_2.

The rest of this paper is organized as follows: Section 2 reviews related work. Section 3 describes an overview of our proposed method. Section 4 describes predicate-argument structure pairs extraction. Section 5 explains co-occurrence statistics calculation between predicate-argument structures using an association rule mining, and Section 6 describes argument alignment based on case frames. Section 7 reports on our experiments.

2 Related Work

We describe manually constructed resources for event relations, and then explain automatic acquisition methods from a corpus.

2.1 Manually Constructed Resource

Singh and Williams constructed a common sense knowledge base concerned with ordinary human activity (Singh and Williams, 2003). The knowledge base consists of 80,000 propositions with 415,000 temporal and atemporal links between propositions. Espinosa and Lieberman proposed an EventNet, a toolkit for inferring temporal relations between commonsense events from the Openmind Commonsense Knowledge Base (Espinosa and Lieberman, 2005).

Recently, Regneri et al. collect natural language descriptions from volunteers over the Internet, and compute a temporal script graph (Regneri et al., 2010). They collected 493 event sequence descriptions for the 22 scenarios such as "eating in a fast-food restaurant" using the Amazon Mechanical Turk.

2.2 Automatic Acquisition of Event Relation from Corpus

There are several types in the event relation acquisition. One is the inference rule acquisition. Lin and Pantel extended the distributional hypothesis on words, and calculated two paths in a dependency tree (Lin and Pantel, 2001). If two paths tend to link the same sets of words, these are regarded as being similar. For example, they calculated the similarity between "X is the author of Y" and "X wrote Y".

Another type is the script-like knowledge acquisition. Chambers and Jurafsky learn narrative schemas, which mean coherent sequences or sets of events, from unlabeled corpora (Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009). This method extracts two events that share a participant, called a *protagonist*. Since these methods rely on the coreference analysis result, they are hard to be applied to languages where omitted arguments or zero anaphora are often utilized.

Kasch and Oates proposed a method for extracting script-like structures from collections of Web documents (Kasch and Oates, 2010). Their method is topic-driven, and the experiment was performed on only one situation *eating at a restaurant*.

There is some work for acquiring two related events taking argument sharing approach (Torisawa, 2006; Abe et al., 2008). Torisawa proposed a method for acquiring inference rules with temporal constrains by using verb-verb co-occurrences in Japanese coordinated sentences and verb-noun co-occurrences (Torisawa, 2006). Abe et al. acquire semantic relations between events by coupling the pattern-based relation-oriented approach and the anchor-based argument-oriented approach (Abe et al., 2008). Their method first acquires candidate predicate pairs by exploiting a pattern-based method, and then seeks anchors indicative of the shared argument. If anchors are found, the predicate pair is verified. These methods can acquire only event relations that have a shared argument.

3 Overview of Our Proposed Method

This paper focuses on Japanese, and extracts two strongly-related events in the form as shown in Figure 1. Figure 2 depicts an overview of our proposed method. First, pairs of predicate-argument structures (PAs) that have a dependency relation are extracted from a Web corpus. Then, from a large number of extracted pair of PAs, strongly-related two predicates with their relevant arguments are extracted. Since the meaning of a predicate itself is often ambiguous, the predicate with

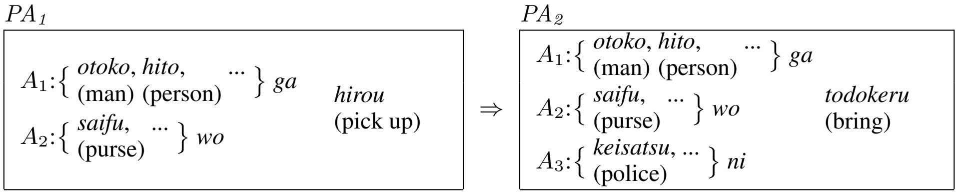

Figure 1: An example of strongly-related events. (*ga* (nom), *wo* (acc), and *ni* (dat) are Japanese case markers.)

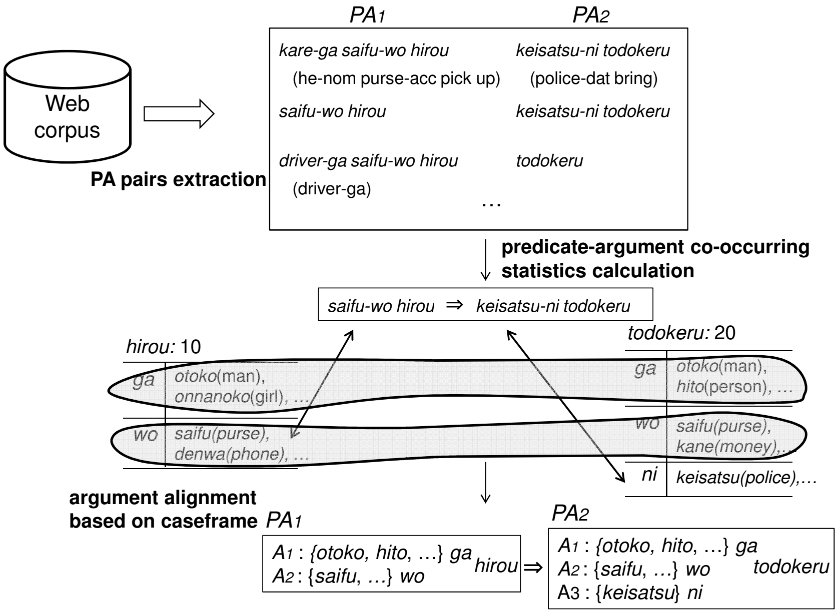

Figure 2: An overview of our proposed method.

their relevant arguments is extracted. For example, whereas the pair between "*hirou* (pick up)" and "*todokeru* (bring)" is not strongly related, the pair between "*saifu-wo* (purse-acc) *hirou*" and "*keisatsu-ni* (police-dat) *todokeru*" is. To extract the predicate with relevant arguments, the pointwise mutual information of the pair of arbitrary PAs is calculated, and the pair whose pointwise mutual information is high is regarded as strongly-related events. We adopt *association rule mining* (Agrawal et al., 1993) for the calculation of co-occurrence statistics between PAs effectively.

Next, the remaining arguments are identified using case frames. For the predicate "*hirou*(pick up)" whose argument takes "*saifu*(purse)-*wo*", what kinds of arguments are taken can be obtained from case frames. As shown in Figure 2, in the case frame 10 of "*hirou*"[2], where the argument *wo* takes "*saihu*", "*denwa*", the argument *ga* takes "*otoko*", "*onnanoko*", and so on. Similarly, in the case frame 20 of "*todokeru*", where the argument *ni* takes "*keisatsu*", the argument *ga* takes "*otoko*", "*onnanoko*", and so on, and the argument *wo* takes "*saihu*" and so on. With the similarity of noun distribution between an argument in PA_1 and one in PA_2, the remaining arguments can be aligned.

[2]Each predicate has several case frames, and case frame 10 of "*hirou*" means 10th case frame for the predicate "*hirou*".

4 Predicate-Argument Structure Pairs Extraction

Strongly-related events appear in the form where they have a dependency relation with a variety of expressions (especially clause relation) in a text. For example, the event "*saifu*(purse)-*wo hirou*(pick up)" and the event "*keisatsu*(police)-*ni todokeru*(bring)" appear as follows:

(2) *saifu-wo hiro-te keisatsu-ni todoke-ta*
purse-acc pick up and police-dat brought
((A man) picked up a purse, and brought it to a police.)

We extract two strongly-related events from a large number of pairs of two PAs that have a dependency relation. From parsing results, a pair of PAs that have a dependency relation is first extracted. The extracted arguments are *ga* (nom), *wo* (acc), and *ni* (dat). If a predicate has an at-

Table 1: Examples of clause relation and predicate-argument structure extraction.

clause relation	example sentence	PA_1	PA_2
sequence	*hachi-ni sa-sare* ***te*** *hareta* (bee-dat)(bitten) (swollen)	*hachi-ni sa-sareru*	*hareru*
cause	*hachi-ni sa-sareta* ***node*** *hareta*	*hachi-ni sa-sareru*	*hareru*
condition	*hachi-ni sa-sareru* ***to*** *hareta*	*hachi-ni sa-sareru*	*hareru*
purpose	*suibun-wo tobasu* ***tame-ni*** *kanetsu-suru* (water-acc) (drain) (heat)	*kanetsu-suru*	*suibun-wo tobasu*
contradiction	*hachi-ni sa-sareta* ***keredo*** *hare-nakatta*	*hachi-ni sa-sareru*	*hareru*
simultaneous	shower-*wo abi* ***nagara*** *ha-wo migaku* (take) (teeth-acc) (brush)	shower-*wo abi*	*ha-wo migaku*

Table 2: Examples of word class and its words.

class	words
77	*hachi* (bee), *ka* (mosquito), $\cdots$
105	dress, *ishou* (cloth), suit, $\cdots$
502	address, *bangou* (number), ID, $\cdots$
956	*juugeki*(shooting), *shuugeki*(attack), $\cdots$
1829	*kenshuu* (training), intern, $\cdots$
1901	*douro* (road), *kokudou* (national highway), $\cdots$

tribute, such as negation, causative, and passive, the attribute is attached to the predicate as a flag. Table 1 shows examples of clause relation and predicate-argument structure extraction.

We consider PA pairs that occur with a clause relation sequence as standard. In the case of clause relation purpose, PA pairs occur in the following form: PA_2 *tame-ni* PA_1, and so PA_1 and PA_2 are transposed. In the case of the clause relation contradiction, the negation flag in the predicate of PA_2 is reversed.

Argument Generalization

An argument is generalized to its word class so as to alleviate the problem of data sparseness. As a word class, a large-scale clustering result of verb-noun dependency relations (Kazama and Torisawa, 2008) is used. The number of word class is 2,000, and this word class covers one million noun phrases. Table 2 shows examples of a word class and its words.

In pairs of the extracted PAs, the noun n is replaced with the word class $\langle c \rangle$ for which the probability $P(c|n)$ is maximal. For example, "PA_1: *ka*(mosquito) *ni sa-sareru* (bitten), PA_2: *hareru* (swollen)" is changed to "PA_1: $\langle 77 \rangle$ *ni sa-sareru*, PA_2: *hareru*" since "*ka*" belongs to the word class $\langle 77 \rangle$. In the same way, "PA_1: *hachi*(bee) *ni sa-sareru*, PA_2: *hareru*" is changed to "PA_1: $\langle 77 \rangle$ *ni sa-sareru*, PA_2: *hareru*", and thus, these two PAs can be identical.

5 Co-occurrence Statistics Calculation between Predicate-Argument Structures

Given a lot of PAs, as extracted in Section 4, the co-occurrence statistics between PAs is calculated. Since the number of pairs of arbitrary PAs is enormous, a question that arises is how to obtain related PAs effectively. To solve this problem, we adopt *association rule mining* (Agrawal et al., 1993) for the calculation of co-occurrence statistics between PAs. The association rule mining method can efficiently seek candidate items that satisfy specific conditions.

5.1 Association Rule Mining

Association rule mining is a method for discovering significant rules in a large database (Agrawal et al., 1993). This method is originally designed to discover rules such as "a customer who buys diapers tends to buy beer" in customer transactions.

Let $I = I_1, I_2, \cdots, I_m$ be a set of binary attributes, called items. Transaction t is defined as a set of items ($t \subseteq I$), and transaction database T is defined as a set of transactions ($T = t_1, t_2, \cdots, t_n$).

A *rule* is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \phi$. This signifies "if X occurs, Y tends to occur". The set of items X and Y are called antecedent (left-hand side, lhs) and consequent (right-hand side, rhs) of the rule respectively. For every rule, the following three measures are defined:

$$support(X \Rightarrow Y) = \frac{C(X \cup Y)}{|T|} \quad (1)$$

$$confidence(X \Rightarrow Y) = \frac{support(X \Rightarrow Y)}{support(X)} \quad (2)$$

$$lift(X \Rightarrow Y) = \frac{confidence(X \Rightarrow Y)}{support(Y)}, \quad (3)$$

Table 3: Examples of transaction data. (One line represents a transaction.)

PA_1		PA_2	
arguments	predicate	arguments	predicate
saifu(purse)-*wo*	*hirou* (pick up)	*keisatsu*(police)-*ni*	*todokeru* (bring)
kare(he)-*ga*, *saifu-wo*	*hirou*	*keisatsu-ni*	*todokeru*
saifu-wo	*hirou*		*todokeru*
	hirou	*keisatsu-ni*	*todokeru*
	...		
saifu-wo	*hirou*		*tewatasu* (hand)
saifu-wo	*hirou*	*kare*(he)-*ni*	*tewatasu*
otoko(man)-*ga*, *saifu-wo*	*hirou*		*tewatasu*
	...		

where $C(X)$ represents the number of transactions containing the item X.

The *support* is defined as the fraction formed the number of transactions that contain the itemset X and the total number of transactions in the database. The *confidence* is defined as the fraction formed from the transactions that contain $X \cup Y$ and the transactions that contain X. The *lift* corresponds to pointwise mutual information between X and Y.

Apriori algorithm (Borgelt and Kruse, 2002) is one of the well-known implementations for association rule mining. This algorithm exploits the observation that no superset of an infrequent itemset can be frequent, and uses breadth-first search and a tree structure to seek candidate items.

The input for Apriori algorithm is transaction data, the minimum support, and minimum confidence, and the algorithm enumerates all rules that satisfy the specified conditions.

5.2 Apriori Algorithm Application to Co-occurrence Calculation

The Apriori algorithm is applied to the calculation of co-occurrence statistics between PAs. An item introduced in Section 5.1 corresponds to a predicate or an argument, and a transaction is obtained from a pair of PAs. Examples of transaction data are shown in Table 3.

Since the rules we want to extract are supposed to satisfy the following conditions:

- X (left-hand side) consists of a predicate of PA_1, and zero or more arguments in PA_1
- Y (right-hand side) consists of a predicate of PA_2, and zero or more arguments in PA_2,

all the rules that do not satisfy these conditions are discarded. Among those that do, the rule for which the lift is higher than *lift-min* and less than *lift-max* is adopted. It is well-known that the pointwise mutual information (which corresponds to lift) for which the frequency is low gets extremely high, and thus rules for which the lift is greater than *lift-max* are discarded.

The Apriori algorithm naturally judges which argument is relevant for each predicate pair. For example, from the transaction data shown in Table 3, the following rule is obtained:

1. *saifu-wo hirou* ⇒ *keisatsu-ni todokeru*

2. *saifu-wo hirou* ⇒ *tewatasu*

The first rule implies that for the predicate pair "*hirou*" and "*todokeru*", "*saifu-wo*" for the predicate in PA_1 and "*keisatsu-ni*" for the predicate in PA_2 are relevant. Similarly, the second rule implies that for the predicate pair "*hirou*" and "*tewatasu*", "*saifu-wo*" for the predicate in PA_1 is relevant.

6 Argument Alignment based on Case Frames

As mentioned in Introduction, since an argument is often omitted in the extracted predicate-argument pairs, there is usually a lack of arguments in the extracted rules as described in the previous section. In the following rule, the argument of the *wo* case in PA_1 corresponds to the *wo* case in PA_2, and the argument that includes nouns such as "*otoko*(man)", "*hito*(person)" acts for the *ga* case both in PA_1 and PA_2.

saifu-wo hirou ⇒ *keisatsu-ni todokeru*

Such alignment between arguments can be performed by case frames. The case frames are constructed automatically by clustering similar predicate usages from a raw corpus, and thus each predicate has several case frames. Examples of the case frames are shown in Table 4. When both a

Table 4: Examples of the automatically constructed case frames.

verb	case marker	examples
hirou:1 (pick up)	*ga*	*josei*(lady), *hito*(person), $\cdots$
	wo	taxi, *kuruma*(car), $\cdots$
		$\cdots$
hirou:10 (pick up)	*ga*	*otoko*(man), *onnanoko*(girl), $\cdots$
	wo	*saifu*(purse), *denwa*(phone) $\cdots$
		$\cdots$
todokeru:1 (deliver)	*ga*	staff, *syokuin*(staff), $\cdots$
	wo	*jyohou*(information), news, $\cdots$
		$\cdots$
todokeru:20 (bring)	*ga*	*otoko*(man), *hito*(person), $\cdots$
	wo	*saifu*(purse), *kane*(money), $\cdots$
	ni	*keisatsu*(police), $\cdots$
		$\cdots$

case in cf_1 assigned to PA_1 and a case in cf_2 assigned to PA_2 have a similar distribution of examples, the case in PA_1 and the case in PA_2 can be aligned.

The best combinations of the case frame in both PA_1 and PA_2 and the best alignment of cases are determined as follows:

1. If there is an argument, select case frames corresponding to the argument, otherwise, all case frames are candidates. In the above example, while in PA_1 the case frame 10 is selected according to the argument for the case *wo* ("*saifu*"), in PA_2 the case frame 20 is selected according to the case *ni* ("*keisatsu*").
2. Choose the best case frame pairs that maximize the following score:

$$\operatorname*{argmax}_{cf_1, cf_2} max_{\mathbf{a}} \sum_{a \in \mathbf{a}} sim(arg_1, a(arg_1)) \quad (4)$$

where $\mathbf{a}$ denotes the alignment of case components between PA_1 and PA_2, arg_1 denotes an argument in PA_1, $a(arg_1)$ denotes an argument in PA_2 that aligned with arg_1, and sim denotes the cosine similarity of the case components distribution between arg_1 and $a(arg_1)$. In the example, the alignment between the case *ga* of the case frame 10 in PA_1 and the case *ga* of the case frame 20 in PA_2, and the case *wo* in PA_1 and the case *wo* in PA_2 is performed.

7 Experiments

7.1 Settings

Approximately 100 million Japanese Web pages were used to extract strongly-related events. These pages include 6 billion sentences, containing 100 billion words. Owing to the presence of many duplicate pages on the Web, such as mirror pages, duplicate sentences were discarded. Thus, 1.6 billion sentences containing approximately 25 billion words were acquired. The average number of characters and words in a sentence were 28.3 and 15.6, respectively.

Table 5: Accuracy of extracted rule and the argument alignment.

	correct		incorrect
extracted rule	96(96.0%)		4(4.0%)
argument alignment	correct	incorrect	
	76(79.1%)	20(20.8%)	–

The Web corpus was processed using the Japanese Morphological Analyzer JUMAN[3] and the Japanese parser KNP[4], and pairs of PAs were extracted. The number of extracted PAs was approximately 400 million.

In the application of Apriori algorithm explained in Section 5.2, the minimum support, confidence was set to 1.0×10^{-7}, 1.0×10^{-3} respectively, and *lift-min*, *lift-max* was set to 10, 10,000 respectively.

The case frames were automatically constructed from the Web corpus consisting 1.6G sentences with a method proposed by (Kawahara and Kurohashi, 2006). For 31,000 predicates, case frames were constructed; the average number of case frames of a predicate was 25; the average number of case slots of a case frame was 4.7.

7.2 Result and Discussion

7.2.1 Evaluation of Co-occurrence Statistics Calculation

We acquired approximately 20,000 rules described in Section 5, and evaluated the acquired rules. We chose 100 rules at random, and evaluated whether each is valid. The upper part in Table 5 shows the accuracy, and we found 96 valid rules of the 100, and the accuracy was 0.96. Examples of the extracted rules and its evaluation are shown in Table 6. A major error is the parsing error. In the example (8) in Table 6, the predicate "*ataru*" in PA_1 is correctly a part of function expressions.

7.2.2 Evaluation of Argument Alignment

We chose 96 instances that were judged as correct in the previous section, and calculated the accu-

[3]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html
[4]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp-e.html

Table 6: Examples of acquired rules by the association rule mining method (Section 5).

	PA_1 argument	PA_1 predicate		PA_2 argument	PA_2 predicate	evaluation
(1)	*teiin*(capacity) *ni*	*tassuru*(reach)	⇒		*shimekiru*(close)	correct
(2)	*daigaku*(university) *wo*	*sotsugyo*(graduate)	⇒	*kaisha*(company) *ni*	*shuusyoku*(get a job)	correct
(3)		*tentou*(fall down)	⇒		*kossetsu*(fracture)	correct
(4)		*nominate-sareru*(nominate)	⇒		*jusyo*(win an award)	correct
(5)		*tazuneru*(visit)	⇒	*hanashi*(talk) *wo*	*ukagau*(hear)	correct
(6)		*purezento*(present)	⇒		*yorokoba-reru*(delighted)	correct
(7)		*kekkon*(get married)	⇒	*kodomo*(child)-*ga*	*iru*(have)	correct
(8)	*riyou*(use)-*ni*	*ataru* (at)	⇒	*touroku*(registration)-*ga*	*hitsuyou*(necessary)	incorrect

Table 7: Examples of acquired strongly-related events. (The underlined arguments indicate the one acquired by the association rule mining method. Ids in the left column correspond to ones in Table 6.)

	PA_1 argument	PA_1 predicate		PA_2 argument	PA_2 predicate	evaluation
(1)	A_1:{ *boshuu*, *moushikomi*, ... (invitation) (application) } *ga* A_2:{ *teiin* (capacity) } *ni*	*tassuru* (reach)	⇒	A_1: { *boshuu*, *moushikomi*, ... (invitation) (application) } *wo*	*shimekiru* (close)	correct
(2)	A_1:{ *watashi*, *kodomo*, ... (I) (child) } *ga* A_2:{ *daigaku* (university) } *wo*	*sotsugyo* (graduate)	⇒	A_1:{ *watashi*, *kodomo*, ... (I) (child) } *ga* A_3:{ *kaisha* (company) } *ni*	*shuusyoku* (get a job)	correct
(3)	A_1:{ *musuko*, *kodomo*, ... (son) (child) } *ga*	*tentou* (fall down)	⇒	A_1:{ *musuko*, *kodomo*, ... (son) (child) } *ga*	*kossetsu* (fracture)	correct
(4)	A_1:{ *sakuhin*, ... (product) } *ga* A_2:{ *sho*, *yuushuu-sho*, ... (prize) (grand prix) } *ni*	*nominate-* *sareru* (nominate)	⇒	A_1:{ *sakuhin*, ... (product) } *ga* A_2:{ *sho*, *yuushuu-sho*, ... (prize) (grand prix) } *wo*	*jusyo* (win an award)	correct
(5)	A_1:{ *watashi*, *hito*, ... (I) (person) } *ga* A_2:{ *sensei*, *shachou*, ... (teacher) (chief) } *wo*	*tazuneru* (visit)	⇒	A_1:{ *watashi*, *hito*, ... (I) (person) } *ga* A_2:{ *sensei*, *shachou*, ... (teacher) (chief) } *ni* A_3:{ *hanashi*, ... (talk) } *wo*	*ukagau* (hear)	correct
(6)	A_1:{ *kanojo*, *josei*, ... (she) (lady) } *ga* A_2:{ *shouhin*, *hana*, ... (goods) (flower) } *wo*	*purezento* (present)	⇒	A_2:{ *shouhin*, *hana*, ... (goods) (flower) } *ga* A_1:{ *kanojo*, *josei*, ... (she) (lady) } *ni*	*yorokoba-* *reru* (delighted)	incorrect
(7)	A_1:{ *kodomo*, ... (child) } *ga*	*kekkon* (get married)	⇒	A_1:{ *kodomo*, ... (child) } *ga*	*iru* (have)	incorrect

racy of the argument alignment. The bottom part in Table 5 shows the accuracy, and we found 76 of 94 were valid, and the accuracy was 0.791. Table 7 shows examples of acquired strongly-related events. A major error is that the case component distribution between two cases in a PA is very similar. In the example (6), the alignment shown in Figure 3 is correct. This error was caused by the fact that the case *ga* and the case *ni* in PA_1 and the case *ga* and the case *ni* in PA_2 include nouns representing an agent.

Another error is that some constructed case frames do not have an indispensable case slot. In the example (7), the alignment shown in Figure 4 is correct. This error is due to the fact that the assigned case frame to PA_2 does not have the *ni* case. To cope with this problem, we are planning to increase the size of Web corpus for the case frames compilation.

7.2.3 Comparison with Coreference-based Method

Our method was compared with the coreference-based method (Chambers and Jurafsky, 2008). Since the accuracy of coreference resolution is not high (Sasano et al. report an F-score of approximately 0.75 in a newspaper domain (Sasano et al., 2007)), if a noun appears twice in a Web page, and it fills a syntactic relation of the predicate w and the predicate v, the noun is regarded as a corefer-

A_1:{ *watashi*, *hito*, ... (I) (person) } *ga*
A_2:{ *shouhin*, *hana*, ... (goods) (flower) } *wo* *purezento* (present)
A_3:{ *kanojo*, *josei*, ... (she) (lady) } *ni*
⇒
A_2:{ *shouhin*, *hana*, ... (goods) (flower) } *ga* *yorokoba-reru* (delighted)
A_3:{ *kanojo*, *josei*, ... (she) (lady) } *ni*

Figure 3: The correct alignment of (6) in Table 7.

A_2:{ *watashi*, *hito*, ... (I) (person) } *ga* *kekkon* (get married)
⇒
A_2:{ *watashi*, *hito*, ... (I) (person) } *ni* *iru* (have)
A_1:{ *kodomo*, ... (child) } *ga*

Figure 4: The correct alignment of (7) in Table 7.

Table 8: Comparison of our method with the coreference-based method. (The covered ratio is the fraction formed the number of the acquired noun by the coreference-based method and the number of the nouns in the aligned argument by our method.)

case in PA_1	case in PA_2	covered ratio	
ga	*ga*	0.163	(3,768 / 23,180)
ga	*wo*	0.282	(549 / 1,944)
ga	*ni*	0.176	(474 / 2,689)
wo	*ga*	0.272	(753 / 2,764)
wo	*wo*	0.483	(7,106 / 14,713)
wo	*ni*	0.321	(1,054 / 3,284)
ni	*ga*	0.163	(344 / 2,113)
ni	*wo*	0.338	(1,042 / 3,086)
ni	*ni*	0.282	(549 / 1,944)

ence, following the method proposed by (Abe et al., 2008). The PMI score was calculated as follows:

$$pmi(e(w,d), e(v,g)) = \log \frac{P(e(w,d), e(v,g))}{P(e(w,d))P(e(v,g))} \quad (5)$$

where $e(w,d)$ is the verb/dependency pair w and d, and d and g have the coreferent relation.

In our acquired rules, we examined whether the k-most frequent noun in the aligned argument can be covered by the coreference-based method. In our experiment, k was set to be 5. The result is shown in Table 8. The number was classified according to the case in PA_1 and in PA_2. We found that most of the nouns in aligned argument cannot be acquired by the coreference-based method. Especially, the covered ratio of the pair of the case *ga* in PA_1 and the case *ga* in PA_2 was relatively low, which often corresponds to *agent*. In Japanese, since an *agent* is often omitted, it is hard to be acquired by the coreference-based method. However, our method can identify its use by using case frames.

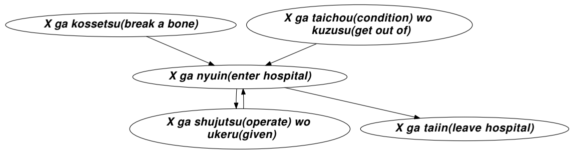

Figure 5: Network structure between events concerned with "enter hospital". (X: {*kodomo*(child), *musume*(daughter), $\cdots$})

7.2.4 Event Network Structure

Figure 5 is an example of a network structure between events concerned with "enter hospital", which is constructed from strongly-related events obtained by our proposed method. Anchor/coreference-based method cannot acquire the argument "*taichou*(condition)-*wo*" that presents in one node (which means this argument is shared by no events). In contrast, our proposed method can acquire such an argument.

8 Conclusion

This paper proposed a method for automatically acquiring strongly-related events from a large corpus using predicate-argument co-occurring statistics and case frames. Our method first extracted pairs of predicate argument structures that have a dependency relation are extracted from a Web corpus. Then, two events whose pointwise mutual information is high is extracted as strongly-related. We adopt association rule mining for the calculation of co-occurrence statistics between predicate-argument structures effectively. Then, the argument alignment was performed by using case frames.

For future work, since the acquired events include several relations such as temporal relation, causality, and means, we are planning to classify the relations automatically. Acquired event relations would then be utilized in Recognizing Textual Entailment (RTE) and Question Answer (QA) tasks.

References

Shuya Abe, Kentaro Inui, and Yuji Matsumoto. 2008. Two-phased event relation acquisition: coupling the relation-oriented and argument-oriented approaches. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1–8.

Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the ACM-SIGMOD 1993 International Conference on Management of Data (1993)*, pages 207–216.

David Bean and Ellen Riloff. 2004. Unsupervised learning of contextual role knowledge for coreference resolution. In *HLT-NAACL 2004: Main Proceedings*, pages 297–304.

Christian Borgelt and Rudolf Kruse. 2002. Induction of association rules: Apriori implementation. In *Proceedings of 15th Conference on Computational Statistics*, pages 395–400.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610.

Jose Espinosa and Henry Lieberman. 2005. EventNet: Inferring temporal relations between commonsense events. In *Proceedings of the 4th Mexican International Conference on Artificial Intelligence*, pages 61–69.

Matthew Gerber and Joyce Chai. 2010. Beyond NomBank: A study of implicit arguments for nominal predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1583–1592.

Niels Kasch and Tim Oates. 2010. Mining script-like structures from the web. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 34–42.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of the HLT-NAACL2006*, pages 176–183.

Jun'ichi Kazama and Kentaro Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of ACL-08: HLT*, pages 407–415.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

Michaela Regneri, Alexander Koller, and Manfred Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 979–988.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2007. Improving coreference resolution using bridging reference resolution and automatically acquired synonyms. In *Discourse Anaphora and Anaphor Resolution Colloquium*, pages 125–136.

Push Singh and William Williams. 2003. Lifenet: A propositional model of ordinary human activity. In *Proceedings of Workshop on Distributed and Collaborative Knowledge Capture*.

Kentaro Torisawa. 2006. Acquiring inference rules with temporal constraints by using japanese coordinated sentences and noun-verb co-occurrences. In *Proceedings of Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL06)*, pages 57–64.