Proceedings of the Federated Conference on Computer Science and Information Systems pp. 339–345
DOI: 10.15439/2017F419
ISSN 2300-5963 ACSIS, Vol. 11

Event Relation Acquisition Using Dependency Patterns and Confidence-Weighted Co-occurrence Statistics

Shohei Higashiyama†, Kunihiko Sadamasa, Takashi Onishi
NEC Corporation
Kanagawa, Japan.
Email: shohei.higashiyama@nict.go.jp, {k-sadamasa@az, t-onishi@bq}.jp.nec.com

Yotaro Watanabe
PKSHA Technology Inc.
Tokyo, Japan.
Email: y_watanabe@pkshatech.com

***Abstract*—Event relation knowledge is important for deep language understanding and inference. Previous work has established automatic acquisition methods of event relations that focus on common sense knowledge acquisition from large-scale unlabeled corpus. However, in the case of domain-specific knowledge acquisition, such a method can not acquire much knowledge due to the limited amount of available knowledge sources. We propose an coverage-oriented acquisition method of event relations. The proposed method utilizes various patterns of dependency structures co-occurring with event relations than the existing method relying only on direct dependency relations between events. Experimental results show that the proposed method can acquire a larger amount of positive relation instances while keeping higher precision compared with the existing method and the proposed method also performs well for small sizes of corpora.**

I. Introduction

SEMANTIC relations between events are important knowledge for various NLP applications that require deep language understanding and inference, such as question answering and future scenario planning. For example, happens-before relation between events (e.g., *get the flu* causes *have a fever*) is required to predict future events from observed events, and entailment relation (e.g., *send a mail to someone* entails *contact someone*) is crucial to recognize similarity of events written with a different surface or in a different abstraction level of expressions.

Many methods have been developed to acquire event relations automatically from unlabeled corpus [1], [2], [3], [4], [5], [6], [7]. Knowledge acquisition methods can be evaluated in terms of accuracy and coverage, and both measures affect performance of downstream applications. In the case of acquiring domain-specific knowledge, we believe that it is important to acquire knowledge with high coverage rather than high accuracy, since accuracy-oriented methods would not acquire much knowledge from the limited amount of available domain corpus. Although coverage-oriented methods extract more incorrect knowledge, eliminating incorrect knowledge from candidates is much easier than creating knowledge not acquired.

We categorize unsupervised or semi-supervised acquisition methods of event relations into the following two types on the basis of how they extract event pairs that correspond to candidate event relations. Methods of the first type [1], [6], [7] extract event pairs from a sentence. They acquire event relations written in a sentence by using syntactic information of the sentence (e.g. dependency relations) or lexical clues indicating clause relations (e.g. expressions such as "because" and "after"). Methods of the second type [2], [3], [4], [5] extract event pairs from sentences in one or more documents. They acquire event relations whose events distantly occur in documents, by using distributional similarities of events or lexical clues indicating sentence relations (e.g. expressions such as "therefore" and "consequently") [1]. Methods of both types usually filter or rank extracted candidates using association measures among predicates and arguments composing a event pair. Note that methods of both types can acquire different relation instances and can be used complementarily [2]. In this work, we focus on event relation acquisition from a sentence.

The existing methods that target a sentence [1], [6], [7] aim to acquire common sense knowledge from large-scale knowledge sources with high accuracy. The methods relying on lexical clues [1], [7] can not acquire relation instances which explicitly occur without lexical expressions indicating event relations. In contrast, Shibata and Kurohashi [6] proposed a method relying almost only on syntax information to extract candidates of happens-before-like relations. Their method extracts event pairs that have dependency relation and ranks those pairs by using pointwise mutual information (PMI) between two events, which measures the degree of co-occurrence. However, their method can not acquire event relations which do not have direct dependency relation.

† Present affiliation is National Institute of Infomation and Communications Technology, Kyoto, Japan.

[1] Some lexical clues, such as discourse connectives, are in common used to detect event relations occurring in a sentence and ones occurring in two sentences.

[2] There is also research that acquires both relation instances occurring in a sentence and ones occurring in two sentences, such as the work by Do et al. [3].

In this work, we propose a method that acquires event relations with high coverage. We introduce various dependency patterns into the calculation approach of association between events by Shibata and Kurohashi. The main differences to that work are as follows:

- Our method detects various dependency patterns between related events and uses the acquired patterns to extract candidate event relations.
- Our method measures the strength of association between two events on the basis of our co-occurrence statistics, namely, the weighted association score. The score is weighted by the confidence of dependency patterns in order to rank instances effectively and obtain high precision.

We performed experiments on Japanese corpora and compared the proposed method with the baseline method that corresponds to the method by Shibata and Kurohashi. The results show that our method efficiently acquires a larger amount of positive relation instances. In addition, our method suppresses decrease of precision against decrease of corpus size and acquires reliable relation instances efficiently from limited sizes of corpora.

II. Related Work

Over the past two decades, many efforts have been focused on automatic acquisition of event relations such as entailment and causality. In particular, unsupervised or semi-supervised methods that target unlabeled corpora have been actively researched.

Torisawa [7], Abe et al. [1], and Shibata and Kurohashi [6] proposed acquisition methods that extract two events co-occurring in a sentence. Torisawa [7] extracts two predicates co-occurring in coordinated sentences to acquire happens-before-like relations. Using a bootstrapping strategy, Abe et al. [1] extract lexico-syntactic patterns co-occurring with given seed relation instances to acquire event relations of the given type, Shibata and Kurohashi [6] use co-occurrence statistics of predicate-argument (PA) pairs, which measures association among all predicates and arguments composing a PA pair, to acquire happens-before-like relations. These methods rely on lexical clues and/or limited syntactic information, and therefore they can acquire limited instances of event relations.

In contrast, acquisition methods that extract two events from multiple sentences have also been proposed. In order to to discover paraphrase-like relations, Lin and Pantel [5] proposed DIRT algorithm, which measures distributional similarity of predicate phrases that are represented by path in dependency tree. Chklovski and Pantel [2] use manually created patterns to classify predicate pairs, which are extracted by DIRT algorithm, into fine-grained relation types such as happens-before and entailment. Hashimoto et al. [4] use distributional similarities between predicates on the basis of common shared arguments to acquire entailment relations. Do et al. [3] use discourse markers and three kinds of associations between predicate-predicate, predicate-argument, and argument-argument to detect causality relations. Do et al. extract event pairs occurring not only in two sentences but also in a sentence. Unlike the work by Do et al., our work targets broader type of relations including entailment and happens-before.

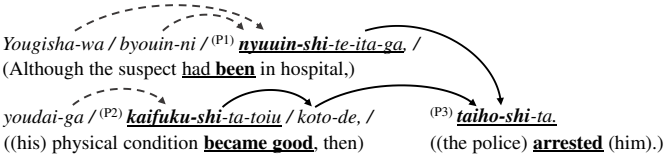

Fig. 1. A dependency tree of a Japanese sentence. Chunks in the sentence are separated by "/". Dependency relations between chunks are denoted by directed edges that are drawn by solid or dotted line. Predicate chunks P1, P2, and P3 are denoted by underline. Predicates are denoted in bold. The smallest subgraph contains P1 and P2, which consists of three solid line edges, indicates the dependency pattern co-occurring with pair of P1 and P2.

In recent years, supervised learning methods have also been applied to learn event relations. Weisman et al. [8] combine various semantic and syntactic features that indicate verb co-occurrence at the sentence, document, and corpus levels to learn entailment relations. Hashimoto et al. [9] use features based on noun relations, such as causality and made-of, between arguments and features based on the association measures of predicates and arguments to learn causality relations. Kloetzer et al. [10] use features based on the transitivity property of entailment to learn entailment relations. These approaches are effective in terms of enlarging existing knowledge base, but they sometimes require a large amount of training instances (e.g., more than tens of thousands of positive instances).

There is a line of research on statistical script models started by Chambers and Jurafsky [11] whereby stereotypical sequences of events (e.g., a visit to a restaurant) is learned. The first model by Chambers and Jurafsky learns statistical scripts involving single participants (e.g., *accuse* X, X *claim*, X *argue*, etc.) on the basis of association between events co-referring to the same protagonist. Chambers and Jurafsky [12] and Pichotta and Mooney [13] extended the first model to handle scripts with multiple protagonists. Recently, embedding-based approaches that can learn script models from large unlabeled corpora have been applied, such as the compositional neural network model by Granroth-Wilding and Clark [14] and the LSTM-based model by Pichotta and Mooney [15]. Unlike our work, these works on script models focus on prediction of missing events in a sequence of events rather than construction of static knowledge.

III. Event Relation Acquisition with High Coverage

As shown in Fig. 1, a dependency tree of a Japanese sentence is expressed as a directed tree where a node represents a chunk and a directed edge represents a dependency relation between chunks [3]. Among all possible combinations

[3]In traditional Japanese dependency parsing, a sentence is divided into chunks, each of which contains one content word and zero or more function words, and then the dependent chunk of each chunk are specified.

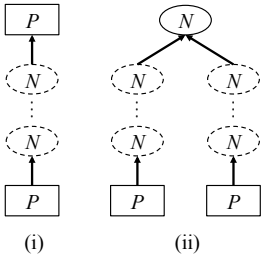

Fig. 2. A dependency pattern between two predicates. Any patte expressed as either (i) a serial or (ii) a parallel dependency structure. S patterns consist of two predicate chunks P and zero or more chunk Parallel patterns consist of two predicate chunks P and one or more ch N. Patterns with the different number of nodes are distinguished from other.

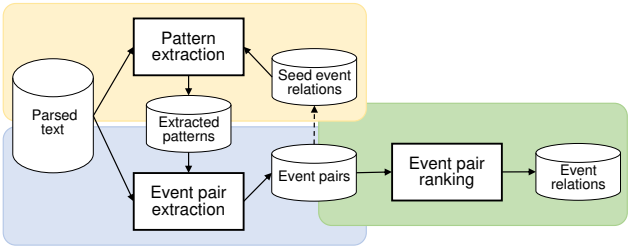

Fig. 3. Framework of proposed system

of two predicates in the sentence, predicate pair ⟨*nyuuin-suru*::*kaifuku-suru*⟩ (⟨be in hospital::become good⟩), which comes from chunk pair of P1 and P2, can be interpreted as a happens-before relation instance.

To extract such relation instances, it is necessary to extract not only direct dependency relations but also various patterns of dependency structures. Here, we assume that every chunk except the root in a parsed sentence has one dependent chunk. In other words, chunks and dependency relations between them constitute a directed tree. Therefore, a dependency relation between any two predicates in a sentence can be expressed as the smallest subgraph that contains the nodes of the two predicate chunks. Since two predicates in a sentence correspond to two leaves in a directed tree, the smallest subgraph that contains the two predicates is equal to either a serial or a parallel dependency pattern as shown in Fig. 2. For example, the dependency pattern that co-occurs with pair of P1 and P2 in Fig. 1 is represented by parallel pattern $\langle P \rightarrow N \leftarrow N \leftarrow P \rangle$, where two P denote the slots of the predicate chunks in interest and the rest N denote the slots of the other chunks in between the predicate chunks. Note that serial pattern with no N nodes corresponds to direct dependency relation and we call other patterns as indirect dependency relations.

In order to improve extraction coverage of various relations instances, we propose an acquisition method that targets both direct and indirect dependency relations between events. An overview of the framework of our system is given in Fig. 3. The system follows three steps below. We assume that input text is dependency-parsed and annotated with dependency relations between chunks, and the parsed text is passed to both pattern acquisition and event pair extraction as input.

1. Pattern extraction
   The system takes parsed text and a small number of seed instances of event relations as input. Then it extracts dependency patterns between events. After that, it calculates the confidence scores of extracted patterns and selects them on the basis of the scores.
2. Event pair extraction
   The system takes parsed text and the extracted patterns as input and then extracts event pairs co-occurring with the patterns.
3. Event pair ranking
   The system calculates the association scores of the extracted event pairs. Event pairs with higher scores are regarded as more reliable event relation instances.

We describe both of the pattern extraction step and the event pair extraction step in Section III-A and the event pair ranking step in Section III-B. We also explain the differences between our method and previous methods. In Section III-C, we describe the weighted association score, which are used in the event pair ranking step, aiming to rank event pairs so that more reliable instances have higher scores.

A. Dependency pattern extraction and event pair extraction

In the pattern extraction and event pair extraction steps, we apply the method by Abe et al. [1]. In this step, unlike the lexico-syntactic patterns in that work, which patterns consisting of word surfaces in directed dependency paths between two predicate chunks, we detect our dependency patterns described above.

a) Pattern extraction step: In the pattern extraction step, our method takes seed relation instances and extracts co-occurrence patterns from input text. Then it calculates confidence scores of patterns so as to enhance the confidence of patterns co-occurring with high-confidence relation instances. From given seed relation instances, confidence score $r_\pi(p)$ for pattern p is calculated as follows:

$$r_\pi(p) = \frac{1}{Z_\pi} \sum_i \mathrm{PMI}(i,p) \cdot r_\iota(i) \ , \qquad (1)$$

where confidence score $r_\iota(i)$ for positive or negative seed instance i is respectively 1 or -1, $\mathrm{PMI}(i,p) = \frac{P(i,p)}{P(i)P(p)}$ is pointwise mutual information between i and p, and Z_π denotes the absolute value of the maximum value of pattern confidence values to normalize the values to $[-1,1]$ [4]. Unlike the work by Abe et al. taking logarithm of the PMI, we define the PMI as above so that confidence of patterns take positive value only when associations with positive instances are stronger than ones with negative instances. The method selects patterns with positive confidence.

[4]We use the normalization similarly to Abe et al. [16] that describes a minor extension of their original work [1].

b) Event pair extraction step: In the event pair extraction step, our method extracts event pairs from input text by using the extracted patterns. At the time, it extracts not only predicates but also arguments depended by the predicates.

In addition, the method also calculates confidence scores r_ι of event pairs defined as follows:

$$r_\iota(i) = \frac{1}{Z_\iota} \sum_p \mathrm{PMI}(i,p) \cdot r_\pi(p) \ , \tag{2}$$

where Z_ι is the coefficient value for normalization defined similarly to Z_π. By using the confidence scores of event relation instances, the method can iterate both extraction steps of patterns and event pairs in a bootstrap manner. These iterations are optional procedures to increase new patterns and event pairs gradually. Then extracted event pairs at the event pair extraction step are passed to the event pair ranking step.

Note that we do not adopt the confidence score of event pairs based on co-occurrence patterns in Eq. (2) unlike Abe et al. Instead, we calculate scores of event pairs on the basis of the direct associations between the events in a similar way to Shibata and Kurohashi.

B. Event pair ranking

In the event pair ranking step, we extend the method by Shibata and Kurohashi.

Our method takes event pairs co-occurring with one or more patterns from the event pair extraction step, and it calculates the PMI between two events as the association score, which we define later. It calculates not only the score of the original event pair but also the scores of any sub-pairs, that is, event pairs comprising two predicates and zero or more arguments of the original event pair. Then the sub-pairs with the highest scores are selected from among any sub-pairs including the original event pair. In the example below, the method also generates event pairs (b), (c), (d) and so on from event pair (a), and it calculates their association scores. Then it selects the pairs with highest scores, which in this case is expected to be pair (b).

(a) ⟨*kodomo-ga kaze-wo hiku*::*netsu-ga deru*⟩ (⟨child catch a cold::have a fever⟩)
(b) ⟨*kaze-wo hiku*::*netsu-ga deru*⟩ (⟨catch a cold::have a fever⟩)
(c) ⟨*kodomo-ga hiku*::*netsu-ga deru*⟩ (⟨child catch::have a fever⟩)
(d) ⟨*hiku*::*deru*⟩ (⟨catch::have⟩)

In order to handle event pairs co-occurring with multiple dependency patterns in a sentence, we define frequency $c(e;s)$ of event e and frequency $c(e,e';s)$ of event pair (e,e') in sentence s so that they take 1 or 0 depending on whether or not it occurs in the sentence, as below:

$$c(e;s) = \begin{cases} 1 & (e \text{ occurs in } s) \\ 0 & (\text{otherwise}) \end{cases}$$

$$c(e,e';s) = \begin{cases} 1 & ((e,e') \text{ co-occurs in } s \\ & \text{with at least one pattern}) \\ 0 & (\text{otherwise}) \ . \end{cases}$$

Consequently, even if an event occurring once in a sentence co-occurs with multiple patterns, the event is not counted redundantly. The association score of event pair (e,e') is calculated from the total frequency $C(e)$ of each event and the total frequency $C(e,e')$ of the event pair in given corpus $\mathcal{C}$, as below:

$$\mathrm{score}(e,e') = \mathrm{PMI}(e,e') = \frac{\frac{C(e,e')}{N}}{\frac{C(e)}{N}\frac{C(e')}{N}} \tag{3}$$

$$C(e) = \sum_{s \in \mathcal{C}} c(e;s) \ , \quad C(e,e') = \sum_{s \in \mathcal{C}} c(e,e';s) \ ,$$

$$N = \sum_e C(e) \ .$$

In addition, we use the discounting factor defined by Pantel and Ravichandran [17] in order to relieve the problem of the PMI being biased towards infrequent elements.

To calculate the PMI of a huge amount of sub-pairs efficiently, we apply Apriori, an association rule mining algorithm, similarly to the work by Shibata and Kurohashi. Association rule mining methods extract subsets of items with strong association from given sets of items as association rules. By pruning unnecessary candidates, Apriori algorithm efficiently calculates several association measures, including the PMI [5], to select strongly-associated rules. We apply the algorithm to event relation acquisition, regarding each event pair (e,e') as a set of predicates and zero or more arguments.

Note that it sometimes happens that extracted instances lack a part of the necessary arguments due to arguments being omitted in text. In addition to the ranking of event pairs, Shibata and Kurohashi make up for lacking arguments of acquired relation instances by using case frames. In this work, we focus on extracting and scoring reliable event relations as the main part of event relation acquisition rather than post-processing to compensate lacking arguments. The extension to compensate arguments in our method remains as future work.

C. Event pair ranking utilizing pattern confidence

In this section, we describe a more sophisticated association score, the weighted association score, between events. The scoring function gives higher scores to event pairs that often co-occur with more reliable patterns and do not often co-occur with more unreliable patterns.

Now, we define weighted frequency $c_w(e,e';s)$ of an event pair in sentence s as

$$c_w(e,e';s) = \begin{cases} \max_{p \in P_{e,e';s}} r_\pi(p) & (\exists p \in P_{e,e';s} \\ & s.t.\ r_\pi(p) > 0) \\ 0 & (\text{otherwise}) \ , \end{cases}$$

where $P_{e,e';s}$ denotes the set of patterns co-occurring with event pair (e,e') in sentence s. Consequence of the normalized value of pattern confidence, weighted frequency $c_w(e,e';s)$ takes at most 1 and does not exceed frequencies of contained events e and e'. Then we define the weighted association score

[5]The association measure corresponding to the PMI is called "lift" in association rule mining.

between event e and e' on the basis of the total weighted frequency C_w of an event pair in given corpus $\mathcal{C}$ as follows:

$$\mathrm{score}_w(e, e') = \frac{\frac{C_w(e,e')}{N}}{\frac{C(e)}{N}\frac{C(e')}{N}} \quad (4)$$

$$C_w(e, e') = \sum_{s \in \mathcal{C}} c_w(e, e'; s) \ . \quad (5)$$

As a result of event pair frequencies being weighted by the confidence of the pattern whose confidence is the highest among co-occurring patterns, the weighted association score provides relatively larger values for event pairs co-occurring with higher-confidence patterns. Therefore it is assumed that the score can rank effectively reliable event pairs.

IV. Experiments

We conducted two experiments to evaluate the proposed method in terms of precision and the amount of acquirable knowledge. First, we compared performance of the baseline method and some versions of the proposed method by using a corpus consisting of 1M documents. The baseline method is the method relying only on direct dependency pattern in the event extraction step of our method, and it corresponds to the method by Shibata and Kurohashi [6]. Second, we also evaluate performance of our method using several smaller sizes of corpora.

A. Experimental Settings

a) Dataset: We use Mainichi newspaper articles (MNA) from 1991 to 2007, which contain about 1.8M documents [7], as input corpus for event relation extraction. In order to compare the performance of methods for a fixed size of input corpus, we use a subset of the corpus in each experiment.

Sentences in the corpus were parsed by CaboCha [18] (version 0.69), a Japanese dependency parser, and then each sentence was divided into chunks and annotated with dependency relations between chunks [8]. From every parsed sentence, we extract verbs as predicates and noun phrases with a case marker *ga* (NOM), *wo* (ACC), or *ni* (DAT) as arguments. However, we eliminated about 20 verbs that are too abstract to be interpreted as meaningful events, such as "*omou* (think)", "*shiru* (know)" and "*motozuku* (be based on)", by choosing among the most frequent verbs in the corpus manually.

b) Parameter settings: For all the compared methods including the baseline and proposed method, we use the below thresholds to filter meaningless instances.

- Threshold of word frequency: Words, either predicate or argument, that occur less than 50 times in a given corpus are cut off. This is because infrequent words are sometimes almost meaningless due to tokenization errors, etc.
- Threshold of event pair frequency: Candidate event relations that occur less than five times in a given corpus are cut off. This is because infrequent elements tend to have a large PMI value but they are not usually reliable.

The proposed method has some additional settings related to pattern extraction.

- Seed relation instances: We manually created five positive instances and five negative instances as seed instances. We chose them from automatically extracted instances from a tiny subset of MNA by the baseline method, which does not require any seed instances.
- Maximum length of patterns: We define the length of a dependency pattern as the length of the corresponding undirected path. On the basis of preliminary experiments with changing the maximum length of extracted patterns, we confirmed that patterns with the larger length tend to have lower confidence. We decided to use at most five-length patterns because negative confidence patterns are extracted when we set five as the maximum length.
- Number of iterations: The pattern extraction and event pair extraction steps can be executed iteratively in a bootstrapping manner. We execute it only once because all possible patterns for each max length constraints were extracted in the first iteration of preliminary experiments.

c) Evaluation Method: Evaluation of the compared methods is done manually by two annotators. Every event pair (e_1, e_2) generated by each method is categorized into three relations by the annotators: happens-before (e_2 often occurs after e_1 occurrs), entailment (e_2 often occurs at the same time as e_1 occurrs), and precondition (e_2 have often occurred before e_1 occurrs).

Some event pairs can not be regarded as positive instances by themselves due to absence of a part of arguments. In case that such event pairs are assumed to have a relation if annotators have compensated suitable additional arguments to them, we allow them as positive instances. In the examples below, pair (a) can be interpreted as a happens-before relation instance by itself. Although pair (b) has a somewhat ambiguous meaning, it can also be regarded as a happens-before relation instance if arguments such as "Website-*ni* (to a Website)" have been compensated for the former predicate. Therefore both examples are expected to be judged as correct.

(a) ⟨*kuuki-ni fureru*::*sanka-suru*⟩ (⟨be exposed to air::get oxidized⟩)

(b) ⟨access-*suru*::page-*wo hiraku*⟩ (⟨access::open a page⟩)

[6]Compared with the baseline method in our experiment, the method by Shibata and Kurohashi additionally utilize case frames for the argument alignment and a word class dictionary for the argument generalization in their work. However, both methods are essentially equivalent in terms of extraction coverage.

[7]We use Mainichi Shimbun CD-ROM (1991-2007) provided by Mainichi Newspapers Co., Ltd. The substantive amount of MNA is actually less because it contains empty documents whose contents have been removed on account of copyright. We eliminated those documents in our experiments.

[8]Although we target dependency relation between chunks in a sentence, our method can be applied to extract dependency relation between words, which is widely used across many languages. However, it should be also examined whether effective patterns to capture event relations are extracted because (typed) word-based dependency patterns between predicates would tend to be longer and more complicated.

TABLE I
PRECISIONS OF EVENT RELATION ACQUISITION FROM THE CORPUS OF 1M DOCUMENTS. PRECISION FOR EACH SECTION AND FOR OVERALL SECTIONS ARE LISTED. DP AND MP INDICATE THE METHOD THAT USE ONLY THE DIRECT PATTERN AND MULTIPLE PATTERNS RESPECTIVELY. MPW IS THE METHOD THAT USE WEIGHTED ASSOCIATION SCORE IN ADDITION TO MULTIPLE PATTERNS.

Method	Precision for each section						Overall
	1-10k	10k-30k	30k-70k	70k-150k	150k-310k	310k-495k	
DP	0.56	0.46	0.18	0.10	–	–	0.231
MP	0.62	0.54	0.52	0.32	0.16	0.10	0.217
MPW	0.72	0.54	0.36	0.24	0.08	0.10	0.167

TABLE II
THE ESTIMATED AMOUNT OF ACQUIRABLE POSITIVE INSTANCES AND THE TOTAL NUMBER OF OUTPUT INSTANCES FROM THE CORPUS OF 1M DOCUMENTS. THE AMOUNT OF POSITIVE INSTANCES INCLUDED IN TOP N INSTANCES ($N = 10\text{K}, 30\text{K}, 70\text{K}, 150\text{K}, 310\text{K}, 495\text{K}$) IS ESTIMATED FROM THE PRECISION IN TABLE I.

Method	No. of positive instances in top N instances						No. of outputs
	∼10k	∼30k	∼70k	∼150k	∼310k	∼495k	
DP	5.6k	14.8k	22.0k	26.4k	–	–	114k
MP	6.2k	17.0k	37.8k	63.4k	89.0k	107.4k	495k
MPW	7.2k	18.0k	32.4k	51.6k	64.4k	82.8k	495k

We used the Cohen's kappa coefficient to measure the inter-annotator agreement, resulting in 0.55 ("moderate" agreement). We adopt each event pair as a positive instance only if two annotators judged it as correct.

B. Experiment 1: Comparison of performance among methods using fixed size of input corpus

In this experiment, we compare the following methods using the subcorpus of MNA consisting of 1M documents. Two versions of our method using multiple dependency patterns, MP and MPW, differ on scoring functions for ranking candidate event relations.

- DP: The baseline method using only direct dependency pattern, which corresponds to the method by Shibata and Kurohashi.
- MP: The proposed method using the ordinal association score in Eq. (3).
- MPW: The proposed method using the weighted association score in Eq. (4).

To estimate precision of each system, we divided relation instances output by each system into sections on the basis of their rank, that is, sections of 1^{st}-10k^{th}, 10k^{th}-30k^{th}, 30k^{th}-70k^{th}, 70k^{th}-150k^{th}, 150k^{th}-310k^{th}, and 310k^{th}-last instances. Then, from each section, 50 relation instances were randomly sampled and judged by annotators. We show the precision for each method for each section in Table I.

The number of output instances from DP is 114k and those of MP and MPW are both 495k (These numbers are also shown in Table II). This result shows that higher ranked instances have higher precision in common among all methods. If we look at the precision of each system, both MP and MPW consistently outperform DP in all sections. In contrast, the overall precisions, which are estimated from all sections, of the proposed methods are lower than that of DP. For the practical purpose of obtaining positive relation instances from among automatically acquired instances by systems, we assume that high rank instances keeping high precision are selected, and then those instances are cleaned by human check to be used for applications. From this point of view, more desirable method should keep higher precision in a wider range of ranked instances, and in that sense the proposed methods are more effective. Besides, in terms of score functions in the proposed methods, MPW performs best in the section of highest-rank instances although MP has the same or higher precision in the rest of the sections. From these results, we confirmed that weighted association score is effective to detect specifically reliable instances but does not maintain robust performance against all acquirable instances.

We also show in Table II the amounts of acquirable positive relation instances from each method, as estimated by the precision in each section and the total numbers of output instances. The results show that MP and MPW can acquire more than three times the amount of positive instances than DP due to use of patterns associated with seed relation instances. Note that, although the two proposed methods only differ on scoring functions, the numbers of acquirable positive instances by the methods are different. This difference, which corresponds to an error rate of about 5% against the total number of outputs, is caused by biases of the random samples.

C. Experiment 2: Validation of performance using different size of input corpus

In order to validate our method perform effectively if only a small size of corpus is available, we perform an evaluation using subcorpora consisting of 500k, 250k, and 100k documents of MNA. We evaluate precision of the top 20% instances for each method, assuming it provides just about the upper bound of precision of each method. The precision of each method and each subcorpus is calculated from 50 random samples similarly to the first experiment.

Table III shows the total number of output instances and the precision of the top 20% instances acquired by each method. Due to the difference between the numbers of output instances by the baseline method and those of the proposed methods, we can not directly compare the precisions between them. However, it is considered to be easier to acquire a larger amount of positive relation instances by the proposed methods because of the increased numbers of output instances compared to those of the baseline method.

Similarly to the first experiment, MPW outperformed MP for all input corpora. The results also show that the precisions of the proposed methods for the smaller size of corpora does not substantially decrease compared with ones for the larger size of corpora. Namely, our methods suppress decrease of precision against decrease of corpus size. Therefore we conclude that our methods can be applied to a limited size of domain corpus for efficient acquisition of reliable relation instances. We plan to acquire domain-specific knowledge by

TABLE III
THE TOTAL NUMBER OF OUTPUT INSTANCES AND PRECISION OF THE TOP 20% INSTANCES ACQUIRED BY EACH METHOD FROM EACH CORPUS OF 500K, 250K, AND 100K DOCUMENTS

	Method	500k-docs	250k-docs	100k-docs
No. of outputs	DP	49.6k	21.7k	5.9k
	MP/MPW	222.1k	101.4k	27.5k
Precision (top 20%)	DP	0.70	0.76	0.66
	MP	0.44	0.46	0.46
	MPW	0.56	0.60	0.48

applying the methods to various domain corpora in future work.

V. CONCLUSION

We have described our method to acquire event relations with high coverage even from a limited size of knowledge sources. We extended the existing baseline method that relies only on direct dependency relation between events and proposed the method that leverages various dependency patterns co-occurring with event relations. We evaluated our method on a general newspaper corpus in Japanese and found that our method can acquire a larger amount of event relations while keeping higher precision compared with the baseline method. The results also show that our method suppresses decrease of precision against decrease of corpus size and it can acquire reliable relation instances efficiently from a limited size of corpus. In future work, we plan to apply the method to various domain corpora and demonstrate the effectiveness of the acquired knowledge for applications such as probabilistic inference.

REFERENCES

[1] S. Abe, K. Inui, and Y. Matsumoto, "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. doi: 10.3115/1599081.1599082 pp. 1–8. [Online]. Available: http://dx.doi.org/10.3115/1599081.1599082

[2] T. Chklovski and P. Pantel, "VerbOcean: Mining the web for fine-grained semantic verb relations," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 4, 2004, pp. 33–40.

[3] Q. X. Do, Y. S. Chan, and D. Roth, "Minimally supervised event causality identification," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011, pp. 294–303.

[4] C. Hashimoto, K. Torisawa, K. Kuroda, S. De Saeger, M. Murata, and J. Kazama, "Large-scale verb entailment acquisition from the web," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics, 2009, pp. 1172–1181.

[5] D. Lin and P. Pantel, "Dirt - discovery of inference rules from text," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 2001. doi: 10.1145/502512.502559 pp. 323–328. [Online]. Available: http://dx.doi.org/10.1145/502512.502559

[6] T. Shibata and S. Kurohashi, "Acquiring strongly-related events using predicate-argument co-occurring statistics and case frames," in *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, 2011, pp. 1028–1036.

[7] K. Torisawa, "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-NAACL)*, 2006. doi: 10.3115/1220835.1220843 pp. 57–64.

[8] H. Weisman, J. Berant, I. Szpektor, and I. Dagan, "Learning verb inference rules from linguistically-motivated evidence," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012, pp. 194–204.

[9] C. Hashimoto, K. Torisawa, J. Kloetzer, M. Sano, I. Varga, J.-H. Oh, and Y. Kidawara, "Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features." in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014, pp. 987–997.

[10] J. Kloetzer, K. Torisawa, C. Hashimoto, and J.-H. Oh, "Large-scale acquisition of entailment pattern pairs by exploiting transitivity," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 1649–1655.

[11] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative event chains," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 94305, 2008, pp. 789–797.

[12] ——, "Unsupervised learning of narrative schemas and their participants," in *Proceedings of Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, 2009. doi: 10.3115/1690219.1690231 pp. 602–610.

[13] K. Pichotta and R. J. Mooney, "Statistical script learning with multi-argument events," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, vol. 14, 2014. doi: 10.3115/v1/e14-1024 pp. 220–229.

[14] M. Granroth-Wilding and S. Clark, "What happens next? event prediction using a compositional neural network model," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, 2016, pp. 2727–2733.

[15] K. Pichotta and R. J. Mooney, "Learning statistical scripts with lstm recurrent neural networks," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, 2016, pp. 2800–2806.

[16] S. Abe, K. Inui, and Y. Matsumoto, "Acquiring event relation knowledge by learning cooccurrence patterns and fertilizing cooccurrence samples (in Japanese)," *Journal of Natural Language Processing*, vol. 16, no. 5, pp. 79–100, 2009. doi: 10.5715/jnlp.16.5_79. [Online]. Available: http://dx.doi.org/10.5715/jnlp.16.5_79

[17] P. Pantel and D. Ravichandran, "Automatically labeling semantic classes," in *Proceedings of Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL)*, vol. 4, 2004, pp. 321–328.

[18] T. Kudo and Y. Matsumoto, "Japanese dependency analysis using cascaded chunking," in *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2002)*, 2002. doi: 10.3115/1118853.1118869 pp. 63–69. [Online]. Available: http://dx.doi.org/10.3115/1118853.1118869