Toward Future Scenario Generation: Extracting Event Causality Exploiting Semantic Relation, Context, and Association Features

Chikara Hashimoto[*] **Kentaro Torisawa**[†] **Julien Kloetzer**[‡] **Motoki Sano**[§]
István Varga[¶] **Jong-Hoon Oh**[‖] **Yutaka Kidawara**[**]
[*†‡§‖**]National Institute of Information and Communications Technology, Kyoto, 619-0289, Japan
[¶]NEC Knowledge Discovery Research Laboratories, Nara, 630-0101, Japan
{[*]ch, [†]torisawa, [‡]julien, [§]msano, [‖]rovellia, [**]kidawara}@nict.go.jp

Abstract

We propose a supervised method of extracting event causalities like *conduct slash-and-burn agriculture→exacerbate desertification* from the web using semantic relation (between nouns), context, and association features. Experiments show that our method outperforms baselines that are based on state-of-the-art methods. We also propose methods of generating *future scenarios* like *conduct slash-and-burn agriculture→exacerbate desertification→increase Asian dust (from China)→asthma gets worse*. Experiments show that we can generate 50,000 scenarios with 68% precision. We also generated a scenario *deforestation continues→global warming worsens→sea temperatures rise→vibrio parahaemolyticus fouls (water)*, which is written in no document in our input web corpus crawled in 2007. But the vibrio risk due to global warming was observed in Baker-Austin et al. (2013). Thus, we "predicted" the future event sequence in a sense.

1 Introduction

The world can be seen as a network of causality where people, organizations, and other kinds of entities causally depend on each other. This network is so huge and complex that it is almost impossible for humans to exhaustively predict the consequences of a given event. Indeed, after the Great East Japan Earthquake in 2011, few expected that it would lead to an enormous trade deficit in Japan due to a sharp increase in energy imports. For effective decision making that carefully considers any form of future risks and chances, we need a system that helps humans do *scenario planning* (Schwartz, 1991), which is a decision-making scheme that examines possible future events and assesses their potential chances and risks. Our ultimate goal is to develop a system that supports scenario planning through generating possible future events using big data, which would contain what Donald Rumsfeld called "unknown unknowns"[1] (Torisawa et al., 2010).

To this end, we propose a supervised method of extracting such event causality as *conduct slash-and-burn agriculture→exacerbate desertification* and use its output to generate *future scenarios* (*scenarios*), which are chains of causality that have been or might be observed in this world like *conduct slash-and-burn agriculture→exacerbate desertification→increase Asian dust (from China)→asthma gets worse*. Note that, in this paper, *A→B* denotes that *A* causes *B*, which means that "*if A happens, the probability of B increases*." Our notion of causality should be interpreted probabilistically rather than logically.

Our method extracts event causality based on three assumptions that are embodied as features of our classifier. First, we assume that two nouns (e.g. *slash-and-burn agriculture* and *desertification*) that take some specific binary semantic relations (e.g. *A* CAUSES *B*) tend to constitute event causality if combined with two predicates (e.g. *conduct* and *exacerbate*). Note that semantic relations are not restricted to those directly relevant to causality like *A* CAUSES *B* but can be those that might seem irrelevant to causality like *A* IS AN INGREDIENT FOR *B* (e.g. *plutonium* and *atomic bomb* as in *plutonium is stolen→atomic bomb is made*). Our underlying intuition is the observation that event causality tends to hold between two entities linked by semantic relations which roughly entail that one entity strongly affects the other. Such semantic relations can be expressed by (otherwise unintuitive) patterns like *A* IS AN INGREDIENT FOR *B*. As such, semantic relations like the MATERIAL relation can also be useful. (See Sec-

[1]http://youtu.be/GiPe1OiKQuk

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 987–997,
Baltimore, Maryland, USA, June 23-25 2014.

tion 3.2.1 for a more intuitive explanation.)

Our second assumption is that there are grammatical contexts in which event causality is more likely to appear. We implement what we consider likely contexts for event causality as context features. For example, a likely context of event causality (underlined) would be: *CO2 levels rose, so climatic anomalies were observed*, while an unlikely context would be: *It remains uncertain whether if the recession is bottomed the declining birth rate is halted*. Useful context information includes the mood of the sentences (e.g., the uncertainty mood expressed by *uncertain* above), which is represented by lexical features (Section 3.2.2).

The last assumption embodied in our association features is that each word of the cause phrase must have a strong association (i.e., PMI, for example) with that of the effect phrase as *slash-and-burn agriculture* and *desertification* in the above example, as in Do et al. (2011).

Our method exploits these features on top of our base features such as nouns and predicates. Experiments using 600 million web pages (Akamine et al., 2010) show that our method outperforms baselines based on state-of-the-art methods (Do et al., 2011; Hashimoto et al., 2012) by more than 19% of average precision.

We require that event causality be *self-contained*, i.e., intelligible as causality without the sentences from which it was extracted. For example, *omit toothbrushing→get a cavity* is self-contained, but *omit toothbrushing→get a girlfriend* is not since this is not intelligible without a context: *He omitted toothbrushing every day and got a girlfriend who was a dental assistant of dental clinic he went to for his cavity*. This is important since future scenarios, which are generated by chaining event causality as described below, must be self-contained, unlike Hashimoto et al. (2012). To make event causality self-contained, we wrote guidelines for manually annotating training/development/test data. Annotators regarded as event causality only phrase pairs that were interpretable as event causality without contexts (i.e., self-contained). From the training data, our method seemed to successfully learn what self-contained event causality is.

Our scenario generation method generates scenarios by chaining extracted event causality; generating $A \rightarrow B \rightarrow C$ from $A \rightarrow B$ and $B \rightarrow C$. The challenge is that many acceptable scenarios are overlooked if we require the joint part of the chain (B above) to be an exact match. To increase the number of acceptable scenarios, our method identifies compatibility w.r.t causality between two phrases by a recently proposed semantic polarity, *excitation* (Hashimoto et al., 2012), which properly relaxes the chaining condition (Section 3.1 describes it). For example, our method can identify the compatibility between *sea temperatures are high* and *sea temperatures rise* to chain *global warming worsens→sea temperatures are high* and *sea temperatures rise→vibrio parahaemolyticus*[2] *fouls (water)*. Accordingly, we generated a scenario *deforestation continues→global warming worsens→sea temperatures rise→vibrio parahaemolyticus fouls (water)*, which is written in no document in our input web corpus that was crawled in 2007, but the vibrio risk due to global warming has actually been observed in the Baltic sea and reported in Baker-Austin et al. (2013). In a sense, we "predicted" the event sequence reported in 2013 by documents written in 2007. Our experiments also show that we generated 50,000 scenarios with 68% precision, which include *conduct terrorist operations→terrorist bombing occurs→cause fatalities and injuries→cause economic losses* and the above "*slash-and-burn agriculture*" scenario (Section 5.2). Neither is written in any document in our input corpus.

In this paper, our target language is Japanese. However, we believe that our ideas and methods are applicable to many languages. Examples are translated into English for ease of explanation. Supplementary notes of this paper are available at `http://khn.nict.go.jp/analysis/member/ch/acl2014-sup.pdf`.

2 Related Work

For **event causality extraction**, clues used by previous methods can roughly be categorized as lexico-syntactic patterns (Abe et al., 2008; Radinsky et al., 2012), words in context (Oh et al., 2013), associations among words (Torisawa, 2006; Riaz and Girju, 2010; Do et al., 2011), and predicate semantics (Hashimoto et al., 2012). Besides features similar to those described above, we propose semantic relation features[3] that include those that are not obviously related to causality. We show that such thorough exploitation of new and existing features leads to high performance.

[2] A bacterium in the sea causing food-poisoning.

[3] Radinsky et al. (2012) and Tanaka et al. (2012) used semantic relations to *generalize* acquired causality instances.

Other clues include shared arguments (Torisawa, 2006; Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009), which we ignore since we target event causality about two distinct entities.

To the best of our knowledge, **future scenario generation** is a new task, although previous works have addressed similar tasks (Radinsky et al., 2012; Radinsky and Horvitz, 2013). Neither involves chaining and restricts themselves to only one event causality step. Besides, the events they predict must be those for which similar events have previously been observed, and their method only applies to news domain.

Some of the scenarios we generated are written on no page in our input web corpus. Similarly, Tsuchida et al. (2011) generated semantic knowledge like causality that is written in no sentence. However, their method cannot combine more than two pieces of knowledge unlike ours, and their target knowledge consists of nouns, but ours consists of verb phrases, which are more informative.

Tanaka et al. (2013)'s web information analysis system provides a *what-happens-if QA* service, which is based on our scenario generation method.

3 Event Causality Extraction Method

This section describes our event causality extraction method. Section 3.1 describes how to extract event causality candidates, and Section 3.2 details our features. Section 3.3 shows how to rank event causality candidates.

3.1 Event Causality Candidate Extraction

We extract the event causality between two events represented by two phrases from single sentences that are dependency parsed.[4] We obtained sentences from 600 million web pages. Each phrase in the event causality must consist of a predicate with an argument position (*template*, hereafter) like *conduct X* and a noun like *slash-and-burn agriculture* that completes *X*. We also require the predicate of the cause phrase to syntactically depend on the effect phrase in the sentence from which the event causality was extracted; we guarantee this by verifying the dependencies of the original sentence. In Japanese, since the temporal order between events is usually determined by precedence in a sentence, we require the cause phrase to precede the effect phrase. For context feature extraction, the event causality candidates are accompanied by the original sentences from which they were extracted.

Excitation We only keep the event causality candidates each phrase of which consists of *excitation templates*, which have been shown to be effective for causality extraction (Hashimoto et al., 2012) and other semantic NLP tasks (Oh et al., 2013; Varga et al., 2013; Kloetzer et al., 2013a). Excitation is a semantic property of templates that classifies them into *excitatory*, *inhibitory*, and *neutral*. Excitatory templates such as *cause X* entail that the function, effect, purpose or role of their argument's referent is activated, enhanced, or manifested, while inhibitory templates such as *lower X* entail that it is deactivated or suppressed. Neutral ones like *proportional to X* belong to neither of them. We collectively call both excitatory and inhibitory templates excitation templates. We acquired 43,697 excitation templates by Hashimoto et al.'s method and the manual annotation of excitation template candidates.[5] We applied the excitation filter to all 272,025,401 event causality candidates from the web and 132,528,706 remained.

After applying additional filters (see Section A in the supplementary notes) including those based on a stop-word list and a causal connective list to remove unlikely event causality candidates that are not removed by the above filter, we finally acquired 2,451,254 event causality candidates.

3.2 Features for Event Causality Classifier

3.2.1 Semantic Relation Features

We hypothesize that two nouns with some particular semantic relations are more likely to constitute event causality. Below we describe the semantic relations that we believe are likely to constitute event causality.

CAUSATION is the causal relation between two entities and is expressed by binary patterns like *A* CAUSES *B*. *Deforestation* and *global warming* might complete the *A* and *B* slots. We manually collected 748 binary patterns for this relation. (See Section B in the supplementary notes for examples of our binary patterns.)

MATERIAL is the relation between a material and a product made of it (e.g. *plutonium* and

[4] We used a Japanese dependency parser called J.DepP (Yoshinaga and Kitsuregawa, 2009), available at `http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/`.

[5] Hashimoto et al.'s method constructs a network of templates based on their co-occurrence in web sentences with a small number of polarity-assigned seed templates and infers the polarity of all the templates in the network by a constraint solver based on the spin model (Takamura et al., 2005).

atomic bomb) and can be expressed by *A* IS MADE OF *B*. Its relation to event causality might seem unclear, but a material can be seen as a "cause" of a product. Indeed materials can participate in event causality with the help of such template pairs as *A is stolen→B is made* as in *plutonium is stolen→atomic bomb is made*. We manually collected 187 binary patterns for this relation.

NECESSITY's patterns include *A* IS NECESSARY FOR *B*, which can be filled with *verbal aptitude* and *ability to think*. Noun pairs with this relation can constitute event causality when combined with template pairs like *improve A→cultivate B*. We collected 257 patterns for this relation.

USE is the relation between means (or instruments) and the purpose for using them. *A* IS USED FOR *B* is a pattern of the relation, which can be filled with *e-mailer* and *exchanges of e-mail messages*. Note that means can be seen as "causing" or "realizing" the purpose of using the means in this relation, and actually event causality can be obtained by incorporating noun pairs of this relation into template pairs like *activate A→conduct B*. 2,178 patterns were collected for this relation.

PREVENTION is the relation expressed by patterns like *A* PREVENTS *B*, which can be filled with *toothbrushing* and *periodontal disease*. This relation is, so to speak, "negative CAUSATION" since the entity denoted by the noun completing the *A* slot makes the entity denoted by the *B* noun NOT realized. Such noun pairs mean event causality by substituting them into template pairs like *omit A→get B*. The number of patterns is 490.

The experiments in Section 5.1.1 show that not only CAUSATION and PREVENTION ("negative CAUSATION") but the other relations are also effective for event causality extraction.

In addition, we invented the EXCITATION relation that is expressed by binary patterns made of excitatory and inhibitory templates (Section 3.1). For instance, we make binary patterns *A* RISES *B* and *A* LOWERS *B* from excitatory template *rise X* and inhibitory template *lower X* respectively. The EXCITATION relation roughly means that *A* activates *B* (excitatory) or suppresses it (inhibitory). We simply add an additional argument position to each template in the 43,697 excitation templates to make binary patterns. We restricted the argument positions (represented by Japanese postpositions) of the *A* slot to either *ha* (topic marker), *ga* (nominative), or *de* (instrumental) and those of the *B* slot to either *ha*, *ga*, *de*, *wo* (accusative), or *ni* (dative), and obtained 55,881 patterns.

SR1: Binary pattern of our semantic relations that co-occurs with two nouns of an event causality candidate in our web corpus.

SR2: Semantic relation types (e.g CAUSATION and ENTAILMENT) of the binary pattern of SR1. EXCITATION is divided into six sub types based on the excitation polarity of the binary patterns, the argument positions, and the existence of causative markers. A CAUSATION pattern, *B* BY *A*, constitutes an independent relation called the BY relation.

Table 1: Semantic relation features.

Moreover, for broader coverage, we acquired binary patterns that entail or are entailed by one of the patterns of the above six semantic relations. Those patterns were acquired from our web corpus by Kloetzer et al. (2013b)'s method, which acquired 185 million entailment pairs with 80% precision from our web corpus and was used for contradiction acquisition (Kloetzer et al., 2013a). We acquired 335,837 patterns by this method. They are *class-dependent patterns*, which have semantic class restrictions on arguments. The semantic classes were obtained from our web corpus based on Kazama and Torisawa (2008). See De Saeger et al. (2009), De Saeger et al. (2011) and Kloetzer et al. (2013a) for more on our patterns. They collectively constitute the ENTAILMENT relation.

Table 1 shows our semantic relation features. To use them, we first make a database that records which noun pairs co-occur with each binary pattern. Then we check a noun pair (the nouns of the cause and effect phrases) for each event causality candidate, and give the candidate all the patterns in the database that co-occur with the noun pair.

3.2.2 Context Features

We believe that contexts exist where event causality candidates are more likely to appear, as described in Section 1. We developed features that capture the characteristics of likely contexts for Japanese event causality (See Section C in the supplementary notes). In a nutshell, they represent a connective (**C1** and **C2** in Section C), the distance between the elements of event causality candidate (**C3** and **C4**), words in context (**C5** to **C8**), the existence of adnominal modifier (**9** to **C10**), and the existence of additional arguments of cause and effect predicates (**C13** to **C20**), among others.

3.2.3 Association Features

These features measure the association strength between *slash-and-burn agriculture* and *deser-*

AC1: The CEA value, the sum of AC2, AC3, and AC4.
AC2: Do et al.'s S_{pp}. This is the association measure between predicates, which is the product of AC5, AC6 and AC7 below. They are calculated from the 132,528,706 event causality candidates in Section 3.1. We omit Do et al.'s $Dist$, which is a constant since we set our window size to one.
AC3: Do et al.'s S_{pa}. This is the association measure between arguments and predicates, which is the sum of AC8 and AC9. They are calculated from the 132,528,706 event causality candidates.
AC4: Do et al.'s S_{aa}, which is PMI between arguments. We obtained it in the same way as Filter 5 in the supplementary notes.
AC5: PMI between predicates.
AC6 / AC7: Do et al.'s max / IDF.
AC8: PMI between a cause noun and an effect predicate.
AC9: PMI between a cause predicate and an effect noun.

Table 2: CEA-based association features.

tification in *conduct slash-and-burn agriculture→exacerbate desertification* for instance and consist of CEA-, Wikipedia-, definition-, and web-based features. **CEA-based features** are based on the Cause Effect Association (CEA) measure of Do et al. (2011). It consists of association measures like PMI between arguments (nouns), between arguments and predicates, and between predicates (Table 2). Do et al. used it (along with discourse relations) to extract event causality. **Wikipedia-based features** are the co-occurrence counts and the PMI values between cause and effect nouns calculated using Wikipedia (as of 2013-Sep-19). We also checked whether an Wikipedia article whose title is a cause (effect) noun contains its effect (cause) noun, as detailed in Section D.1 in the supplementary notes. **Definition-based features**, as detailed in Section D.2 in the supplementary notes, resemble the Wikipedia-based features except that the information source is the definition sentences automatically acquired from our 600 million web pages using the method of Hashimoto et al. (2011). **Web-based features** provide association measures between nouns using various window sizes in the 600 million web pages. See Section D.3 for detail. Web-based association measures were obtained from the same database as **AC4** in Table 2.

3.2.4 Base Features

Base features represent the basic properties of event causality like nouns, templates, and their excitation polarities (See Section E in the supplementary notes). For **B3** and **B4**, 500 semantic classes were obtained from our web corpus using the method of Kazama and Torisawa (2008).

3.3 Event Causality Scoring

Using the above features, a classifier[6] classifies each event causality candidate into causality and non-causality. An event causality candidate is given a causality score $CScore$, which is the SVM score (distance from the hyperplane) that is normalized to $[0, 1]$ by the sigmoid function $\frac{1}{1+e^{-x}}$. Each event causality candidate may be given multiple original sentences, since a phrase pair can appear in multiple sentences, in which case it is given more than one SVM score. For such candidates, we give the largest score and keep only one original sentence that corresponds to the largest score.[7] Original sentences are also used for scenario generation, as described below.

4 Future Scenario Generation Method

Our future scenario generation method creates scenarios by chaining event causalities. A naive approach chains two phrase pairs by exact matching. However, this approach would overlook many acceptable scenarios as discussed in Section 1. For example, *global warming worsens→sea temperatures are high* and *sea temperatures rise→vibrio parahaemolyticus fouls (water)* can be chained to constitute an acceptable scenario, but the joint part is not the same string. Note that the two phrases are not simply paraphrases; temperatures may be rising but remain cold, or they may be decreasing even though they remain high.

What characterizes two phrases that can be the joint part of acceptable scenarios? Although we have no definite answer yet, we *name* it the *causal-compatibility* of two phrases and provide its preliminary characterization based on the excitation polarity. Remember that excitatory templates like *cause X* entail that *X*'s function or effect is activated, but inhibitory templates like *lower X* entail that it is suppressed (Section 3.1). Two phrases are *causally-compatible* if they mention the same entity (typically described by a noun) that is predicated by the templates of the *same excitation polarity*. Indeed, both *X rise* and *X are high* are excitatory and hence *sea temperatures are high* and *sea temperatures rise* are causally-compatible.[8]

[6] We used SVM^{light} with the polynominal kernel ($d = 2$), available at `http://svmlight.joachims.org`.

[7] Future work will exploit other original sentences, as suggested by an anonymous reviewer.

[8] Using other knowledge like verb entailment (Hashimoto et al., 2009) can be helpful too, which is further future work.

Scenarios (scs) generated by chaining causally-compatible phrase pairs are scored by $Score(sc)$, which embodies our assumption that an acceptable scenario consists of plausible event causality pairs:

$$Score(sc) = \prod_{cs \in CAUS(sc)} CScore(cs)$$

where $CAUS(sc)$ is a set of event causality pairs that constitutes sc and cs is a member of $CAUS(sc)$. $CScore(cs)$, which is cs's score, was described in Section 3.3.

Our method optionally applies the following two filters to scenarios for better precision: An **original sentence filter** removes a scenario if two event causality pairs that are chained in it are extracted from original sentences between which no word overlap exists other than words constituting causality pairs. In this case, the two event causality pairs tend to be about different topics and constitute an incoherent scenario. A **common argument filter** removes a scenario if a joint part consists of two templates that share no argument in our ⟨argument, template⟩ database, which is compiled from the syntactic dependency data between arguments and templates extracted from our web corpus. Such a scenario tends to be incoherent too.

5 Experiments

5.1 Event Causality Extraction

Next we describe our experiments on event causality extraction and show **(a)** that most of our features are effective and **(b)** that our method outperforms the baselines based on state-of-the-art methods (Do et al., 2011; Hashimoto et al., 2012). Our method achieved 70% precision at 13% recall; we can extract about 69,700 event causality pairs with 70% precision, as described below.

For the **test data**, we randomly sampled 23,650 examples of ⟨event causality candidate, original sentence⟩ among which 3,645 were positive from 2,451,254 event causality candidates extracted from our web corpus (Section 3.1). For the **development data**, we identically collected 11,711 examples among which 1,898 were positive. These datasets were annotated by three annotators (not the authors), who annotated the event causality candidates without looking at the original sentences. The final label was determined by majority vote. The **training data** were created by the annotators through our preliminary experiments and consists of 112,110 among which 9,657 were positive. The Kappa (Fleiss, 1971) of their judgments was 0.67 (substantial agreement (Landis and Koch, 1977)). These three datasets have no overlap in terms of phrase pairs. About nine man-months were required to prepare the data.

Method	Ave. prec. (%)
Proposed	**46.27**
w/o Context features	45.68
w/o Association features	45.66
w/o Semantic relation features	44.44
Base features only	41.29

Table 3: Ablation tests.

Semantic relations	Ave. prec. (%)
All semantic relations (Proposed)	**46.27**
CAUSATION	45.86
CAUSATION and PREVENTION	45.78
None (w/o Semantic relation features)	44.44

Table 4: Ablation tests on semantic relations.

Our evaluation is based on *average precision*;[9] we believe that it is important to *rank* the plausible event causality candidates higher.

5.1.1 Ablation Tests

We evaluated the features of our method by ablation tests. Table 3 shows the results of removing the semantic relation, the context, and the association features from our method. All the feature types are effective and contribute to the performance gain that was about 5% higher than the **Base features only**. **Proposed** achieved 70% precision at 13% recall. We then estimated that, with the precision rate, we can extract 69,700 event causality pairs from the 2,451,254 event causality candidates, among which the estimated number of positive examples is 377,794.

Next we examined whether the semantic relations that do not seem directly relevant to causality like MATERIAL are effective. Table 4 shows that the performance degraded (46.27 → 45.86) when we only used the CAUSATION binary patterns and their entailing and entailed patterns compared to **Proposed**. Even when adding the PREVENTION ("negative CAUSATION") patterns and their entailing and entailed patterns, the performance was still slightly worse than **Proposed**. The performance was even worse when using no semantic relation ("None" in Table 4). Consequently we conclude that not only semantic relations directly relevant

[9] It is obtained by computing the precision for each point in the ranked list where we find a positive sample and averaging all the precision figures (Manning and Schütze, 1999).

Method	Ave. prec. (%)
w/o Wikipedia-based features	**46.52**
Proposed	46.27
w/o definition-based features	46.21
w/o Web-based features	46.15
w/o CEA-based features	45.80

Table 5: Ablation tests on association features.

Method	Ave. prec. (%)
Proposed	**46.27**
Proposed-CEA	45.80
CEA$_{sup}$	21.77
CEA$_{uns}$	16.57

Table 6: Average precision of our proposed methods and baselines using CEA.

to causality like CAUSATION but also those that seem to lack direct relevance to causality like MATERIAL are somewhat effective.

Finally, Table 5 shows the performance drop by removing the Wikipedia-, definition-, web-, and CEA-based features. The CEA-based features were the most effective, while the Wikipedia-based ones slightly degraded the performance.

5.1.2 Comparison to Baseline Methods

We compared our method and two baselines based on Do et al. (2011): **CEA**$_{uns}$ is an unsupervised method that uses CEA to rank event causality candidates, and **CEA**$_{sup}$ is a supervised method using SVM and the CEA features, whose ranking is based on the SVM scores. The baselines are not complete implementations of Do et al.'s method which uses discourse relations identified based on Lin et al. (2010) and exploits them with CEA within an ILP framework. Nonetheless, we believe that this comparison is informative since CEA can be seen as the main component; they achieved a F1 of 41.7% for extracting causal event relations, but with only CEA they still achieved 38.6%.

Table 6 shows the average precision of the compared methods. **Proposed** is our proposed method. **Proposed-CEA** is **Proposed** without the CEA-features and shows their contribution. **Proposed** is the best and the CEA features slightly contribute to the performance, as **Proposed-CEA** indicates. We observed that **CEA**$_{sup}$ and **CEA**$_{uns}$ performed poorly and tended to favor event causality candidates whose phrase pairs were highly relevant to each other but described the contrasts of events rather than event causality (e.g. *build a slow muscle* and *build a fast muscle*) probably because their

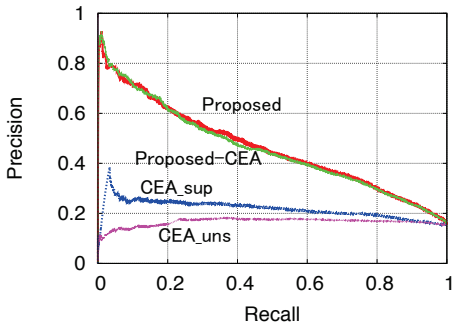

Figure 1: Precision-recall curves of proposed methods and baselines using CEA.

Method	Ave. prec. (%)
Proposed	**49.64**
Cs$_{uns}$	30.38
Cs$_{sup}$	27.49

Table 7: Average precision of our proposed method and baselines using Cs.

main components are PMI values. Figure 1 shows their precision-recall curves.

Next we compared our method with the baselines based on Hashimoto et al. (2012). They developed an automatic excitation template acquisition method that assigns each template an *excitation value* in range $[-1, 1]$ that is positive if the template is excitatory and negative if it is inhibitory. They ranked event causality candidates by $Cs(p_1, p_2) = |s_1| \times |s_2|$, where p_1 and p_2 are the two phrases of event causality candidates, and $|s_1|$ and $|s_2|$ are the absolute excitation values of p_1's and p_2's templates. The baselines are as follows: **Cs**$_{uns}$ is an unsupervised method that uses Cs for ranking, and **Cs**$_{sup}$ is a supervised method using SVM with Cs as the only feature that uses SVM scores for ranking. Note that some event causality candidates were not given excitation values for their templates, since some templates were acquired by manual annotation without Hashimoto et al.'s method. To favor the baselines for fairness, the event causality candidates of the development and test data were restricted to those with excitation values. Since **Cs**$_{sup}$ performed slightly better when using all of the training data in our preliminary experiments, we used all of it.

Table 7 shows the average precision of the compared methods. **Proposed** is our method. Its average precision is different from that in Table 6 due to the difference in test data described above. **Cs**$_{uns}$ and **Cs**$_{sup}$ did not perform well. Many

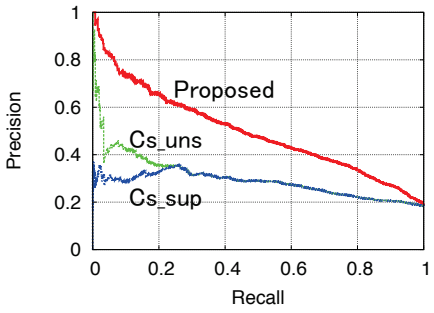

Figure 2: Precision-recall curves of proposed methods and baselines using Cs.

phrase pairs described two events that often happen in parallel but are not event causality (e.g. *reduce the intake of energy* and *increase the energy consumption*) in the highly ranked event causality candidates of **Cs**$_{uns}$ and **Cs**$_{sup}$. Figure 2 shows their precision-recall curves.

Hashimoto et al. (2012) extracted 500,000 event causalities with about 70% precision. However, as described in Section 1, our event causality criteria are different; since they regarded phrase pairs that were not self-contained as event causality (their annotators checked the original sentences of phrase pairs to see if they were event causality), their judgments tended to be more lenient than ours, which explains the performance difference.

In preliminary experiments, since our proposed method's performance degraded when Cs was incorporated, we did not use it in our method.

5.2 Future Scenario Generation

To show that our future scenario generation methods can generate many acceptable scenarios with reasonable precision, we experimentally compared four methods: **Proposed**, our scenario generation method without the two filters, **Proposed+Orig**, our method with the original sentence filter, **Proposed+Orig+Comm**, our method with the original sentence and common argument filters, and **Exact**, a method that chains event causality by exact matching.

Beginning events As the beginning event of a scenario, we extracted nouns that describe social problems (*social problem nouns*, e.g. *deforestation*) from Wikipedia to focus our evaluation on the ability to generate scenarios about them, which is a realistic use-case of scenario generation. We extracted 557 social problem nouns and used the cause phrases of the event causality candidates that consisted of one of the social problem nouns as the scenario's beginning event.

	Two-step	Three-step
Exact	1,000 (44.10)	1,000 (23.50)
Proposed	2,000 (32.25)	2,000 (12.55)
Proposed+Orig	995 (36.28)	602 (17.28)
Proposed+Orig+Comm	708 (38.70)	339 (17.99)

Table 8: Number of scenario samples and their precision (%) in parentheses.

Event causality We applied our event causality extraction method to 2,451,254 candidates (Section 3.1) and culled the top 1,200,000 phrase pairs from them (See Section F in the supplementary notes for examples). Some phrase pairs have the same noun pairs and the same template polarity pairs (e.g. *omit toothbrushing→get a cavity* and *neglect toothbrushing→have a cavity*, where *omit X* and *neglect X* are inhibitory and *get X* and *have X* are excitatory). We removed such phrase pairs except those with the highest $CScore$, and 960,561 phrase pairs remained, from which we generated two- or three-step scenarios that consisted of two or three phrase pairs.

Evaluation samples The numbers of two- and three-step scenarios generated by **Proposed** were 217,836 and 5,288,352, while those of **Exact** were 22,910 and 72,746. We sampled 2,000 from **Proposed**'s two- and three-step scenarios and 1,000 from those of **Exact**. We applied the filters to the sampled scenarios of **Proposed**, and the results were regarded as the sample scenarios of **Proposed+Orig** and **Proposed+Orig+Comm**. Table 8 shows the number and precision of the samples. Note that, for the diversity of the sampled scenarios, our sampling proceeded as follows: **(i)** Randomly sample a beginning event phrase from the generated scenarios. **(ii)** Randomly sample an effect phrase for the beginning event phrase from the scenarios. **(iii)** Regarding the effect phrase as a cause phrase, randomly sample an effect phrase for it, and repeat (iii) up to the specified number of steps (2 or 3). The samples were annotated by three annotators (not the authors), who were instructed to regard a sample as acceptable if each event causality that constitutes it is plausible and the sample as a whole constitutes a single coherent story. Final judgment was made by majority vote. Fleiss' kappa of their judgments was 0.53 (moderate agreement), which is lower than the kappa for the causality judgment. This is probably because

	Two-step	Three-step
Exact	2,085	1,237
Proposed	5,773	0
Proposed+Orig	4,107	0
Proposed+Orig+Comm	3,293	21,153

Table 9: Estimated number of acceptable scenarios with a 70% precision rate.

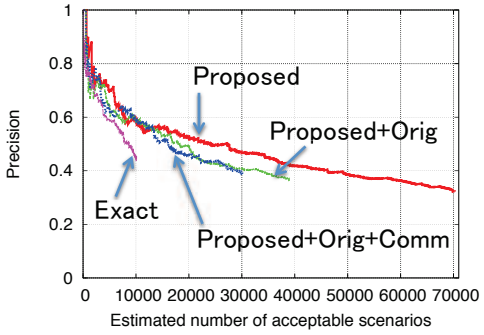

Figure 3: Precision-scenario curves (2-step).

scenario judgment requires careful consideration about various possible futures for which individual annotators tend to draw different conclusions.

Result 1 Table 9 shows the estimated number of acceptable scenarios generated with 70% precision. The estimated number is calculated as the product of the recall at 70% precision and the number of acceptable scenarios in all the generated scenarios, which is estimated by the annotated samples. Figures 3 and 4 show the *precision-scenario curves* for the two- and three-step scenarios, which illustrate how many acceptable scenarios can be generated with what precision. The curve is drawn in the same way as the precision-recall curve except that the X-axis indicates the estimated number of acceptable scenarios. At 70% precision, all of the proposed methods outperformed **Exact** in the two-step setting, and **Proposed+Orig+Comm** outperformed **Exact** in the three-step setting.

Result 2 To evaluate the top-ranked scenarios of **Proposed+Orig+Comm** in the three-step setting with more samples, the annotators labeled 500 samples from the top 50,000 of its output. 341 (68.20%) were acceptable, and the estimated number of acceptable scenarios at a precision rate of 70% and 80% are 26,700 and 5,200 (See Section H in the supplementary notes). The "*terrorist operations*" scenario and the "*slash-and-burn agriculture*" scenario in Section 1 were ranked 16,386th

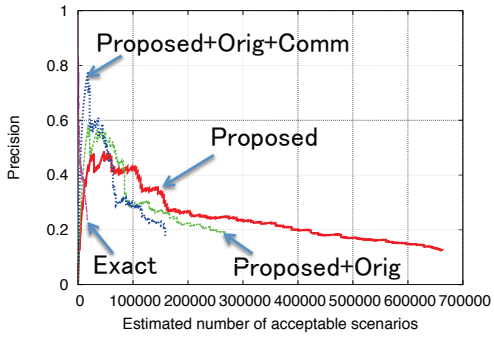

Figure 4: Precision-scenario curves (3-step).

and 21,968th. Next we examined how many of the top 50,000 scenarios were acceptable and *non-trivial*, i.e., found in no page in our input web corpus, using the 341 acceptable samples. A scenario was regarded as non-trivial if its nouns co-occur in no page of the corpus. 22 among the 341 samples were non-trivial. Accordingly, we estimate that we can generate 2,200 ($\frac{50{,}000\times 22}{500}$) acceptable and non-trivial scenarios from the top 50,000. (See Section G in the supplementary notes for examples of the generated scenarios.)

Discussion Scenario *deforestation continues→global warming worsens→sea temperatures rise→vibrio parahaemolyticus fouls (water)* was generated by **Proposed+Orig+Comm**. It is written in no page in our input web corpus, which was crawled in 2007.[10] But we did find a paper Baker-Austin et al. (2013) that observed the emerging vibrio risk in the Baltic sea due to global warming. In a sense, we "predicted" an event observed in 2013 from documents written in 2007, although the scenario was ranked as low as 240,738th.

6 Conclusion

We proposed a supervised method for event causality extraction that exploits semantic relation, context, and association features. We also proposed methods for our new task, future scenario generation. The methods chain event causality by causal-compatibility. We generated non-trivial scenarios with reasonable precision, and "predicted" future events from web documents. Increasing their rank is future work.

[10]The corpus has pages where *global warming*, *sea temperatures*, and *vibrio parahaemolyticus* happen to co-occur. But they are either diaries where the three words appear separately in different topics or lists of arbitrary words.

References

Shuya Abe, Kentaro Inui, and Yuji Matsumoto. 2008. Two-phrased 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.

Susumu Akamine, Daisuke Kawahara, Yoshikiyo Kato, Tetsuji Nakagawa, Yutaka I. Leon-Suematsu, Takuya Kawada, Kentaro Inui, Sadao Kurohashi, and Yutaka Kidawara. 2010. Organizing information on the web to support user judgments on information credibility. In *Proceedings of 2010 4th International Universal Communication Symposium Proceedings (IUCS 2010)*, pages 122–129.

Craig Baker-Austin, Joaquin A. Trinanes, Nick G. H. Taylor, Rachel Hartnell, Anja Siitonen, and Jaime Martinez-Urtaza. 2013. Emerging vibrio risk at high latitudes in response to ocean warming. *Nature Climate Change*, 3:73–77.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 789–797.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP (ACL-IJCNLP 2009)*, pages 602–610.

Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata. 2009. Large scale relation acquisition using class dependent patterns. In *Proceedings of the IEEE International Conference on Data Mining (ICDM 2009)*, pages 764–769.

Stijn De Saeger, Kentaro Torisawa, Masaaki Tsuchida, Jun'ichi Kazama, Chikara Hashimoto, Ichiro Yamada, Jong-Hoon Oh, István Varga, and Yulan Yan. 2011. Relation acquisition using word classes and partial patterns. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 825–835.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 294–303.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Masaki Murata, and Jun'ichi Kazama. 2009. Large-scale verb entailment acquisition from the web. In *Proceedings of EMNLP 2009: Conference on Empirical Methods in Natural Language Processing*, pages 1172–1181.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jun'ichi Kazama, and Sadao Kurohashi. 2011. Extracting paraphrases from definition sentences on the web. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1087–1097.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of EMNLP-CoNLL 2012: Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*, pages 619–630.

Jun'ichi Kazama and Kentaro Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 407–415.

Julien Kloetzer, Stijn De Saeger, Kentaro Torisawa, Chikara Hashimoto, Jong-Hoon Oh, and Kiyonori Ohtake. 2013a. Two-stage method for large-scale acquisition of contradiction pattern pairs using entailment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 693–703.

Julien Kloetzer, Kentaro Torisawa, Stijn De Saeger, Motoki Sano, Chikara Hashimoto, and Jun Gotoh. 2013b. Large-scale acquisition of entailment pattern pairs. In *Information Processing Society of Japan (IPSJ) Kansai-Branch Convention 2013*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A pdtb-styled end-to-end discourse parser. Technical report, School of Computing, National University of Singapore.

Chris Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. 2013. Why-question answering using intra- and inter-sentential causal relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1733–1743.

Kira Radinsky and Eric Horvitz. 2013. Mining the web to predict future events. In *Proceedings of Sixth ACM International Conference on Web Search and Data Mining (WSDM 2013)*, pages 255–264.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of International World Wide Web Conference 2012 (WWW 2012)*, pages 909–918.

Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *2010 IEEE Fourth International Conference on Semantic Computing*, pages 361–368.

Peter Schwartz. 1991. *The Art of the Long View*. Doubleday.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientation of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 133–140.

Shohei Tanaka, Naoaki Okazaki, and Mitsuru Ishizuka. 2012. Acquiring and generalizing causal inference rules from deverbal noun constructions. In *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1209–1218.

Masahiro Tanaka, Stijn De Saeger, Kiyonori Ohtake, Chikara Hashimoto, Makoto Hijiya, Hideaki Fujii, and Kentaro Torisawa. 2013. WISDOM2013: A large-scale web information analysis system. In *Companion Volume of the Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013) (Demo Track)*, pages 45–48.

Kentaro Torisawa, Stijn de Saeger, Jun'ichi Kazama, Asuka Sumida, Daisuke Noguchi, Yasunari Kakizawa, Masaki Murata, Kow Kuroda, and Ichiro Yamada. 2010. Organizing the web's information explosion to discover unknown unknowns. *New Generation Computing (Special Issue on Information Explosion)*, 28(3):217–236.

Kentaro Torisawa. 2006. Acquiring inference rules with temporal constraints by using japanese coordinated sentences and noun-verb co-occurrences. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT-NAACL2006)*, pages 57–64.

Masaaki Tsuchida, Kentaro Torisawa, Stijn De Saeger, Jong Hoon Oh, Jun'ichi Kazama, Chikara Hashimoto, and Hayato Ohwada. 2011. Toward finding semantic relations not written in a single sentence: An inference method using auto-discovered rules. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 902–910.

István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. 2013. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1619–1629.

Naoki Yoshinaga and Masaru Kitsuregawa. 2009. Polynomial to linear: Efficient classification with conjunctive features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 542–1551.