Generating Event Causality Hypotheses through Semantic Relations

Chikara Hashimoto*, Kentaro Torisawa†, Julien Kloetzer‡, Jong-Hoon Oh§
National Institute of Information and Communications Technology, Kyoto, 619-0289, Japan
{*ch, †torisawa, ‡julien, §rovellia}@nict.go.jp

Abstract

Event causality knowledge is indispensable for intelligent natural language understanding. The problem is that any method for extracting event causalities from text is insufficient; it is likely that some event causalities that we can recognize in this world are not written in a corpus, no matter its size. We propose a method of hypothesizing *unseen* event causalities from known event causalities extracted from the web by the semantic relations between nouns. For example, our method can hypothesize *deploy a security camera→avoid crimes* from *deploy a mosquito net→avoid malaria* through semantic relation *A* PREVENTS *B*. Our experiments show that, from 2.4 million event causalities extracted from the web, our method generated more than 300,000 hypotheses, which were not in the input, with 70% precision. We also show that our method outperforms a state-of-the-art hypothesis generation method.

1 Introduction

Event causality knowledge, e.g., *deploy a mosquito net→avoid malaria*, enhances natural language understanding systems like future event prediction (Radinsky, Davidovich, and Markovitch 2012), why-question answering (Oh et al. 2013), and future scenario generation (Hashimoto et al. 2014). In this paper, *A→B* represents event causality that basically means that "*if A happens, the probability of B increases*." We discuss our position on causality below.

Many methods have been proposed that *extract* event causalities from corpora (Torisawa 2006; Abe, Inui, and Matsumoto 2008; Chambers and Jurafsky 2008; 2009; Riaz and Girju 2010; Do, Chan, and Roth 2011; Radinsky, Davidovich, and Markovitch 2012; Hashimoto et al. 2012; 2014). However, methods that *hypothesize* plausible event causalities that are not written in corpora have not been fully explored so far, even though human beings can do it easily.

In this paper, we propose a method of generating plausible *event causality hypotheses* (*hypotheses*, in short) from other event causalities extracted from the web. We assume that if a noun pair bears a semantic relation and constitutes plausible event causality, other noun pairs of the same semantic relation tend to constitute a plausible event causality, too.

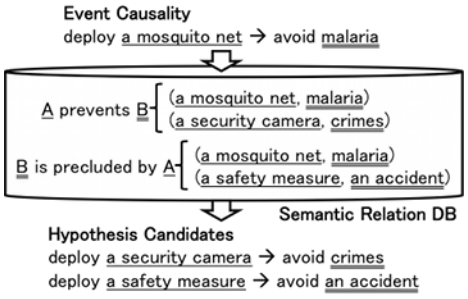

Figure 1: Hypothesis candidate generation.

Take noun pair *a mosquito net* and *malaria*, which bears the *A* PREVENTS *B* relation, for example. The noun pair constitutes a plausible event causality, e.g., *deploy a mosquito net→avoid malaria*. Other noun pairs of the *A* PREVENTS *B* relation, e.g., *a security camera* and *crimes*, also constitute plausible event causalities, e.g., *deploy a security camera→avoid crimes.* Thus, our method replaces a noun pair (the nouns of the cause and effect phrases) of a given event causality with other noun pairs of the same semantic relation to generate event causality hypothesis candidates (Figure 1).

Hypothesis candidates are ranked by our hypothesis classifier, which checks their plausibility as event causalities. Then we train it using labeled data and features that have been developed for event causality extraction tasks (Hashimoto et al. 2014) (Section 2.2).

In short, our method conducts a generate-and-test search for plausible event causality hypotheses; generating a large number of hypothesis candidates using semantic relations and event causalities extracted from the web, and testing them using the hypothesis classifier.

Our experiments show that, using about 2.4 million event causalities extracted from the web by Hashimoto et al. (2014)'s method as a source, our method generated a large number of *unseen* plausible hypotheses (with 70% precision) that were not included in the source event causalities: 347,093 hypotheses whose noun pairs were not in the source (Section 3.1) and 302,350 hypotheses whose phrase pairs were not in the source (Section 3.2). These results indicate that our method can *synthesize* a large volume of unseen yet plausible event causality knowledge from many

pieces of knowledge fragments, i.e., known event causality knowledge and noun pairs bearing specific semantic relations that are scattered across the web. Our method also greatly outperforms a state-of-the-art hypothesis generation method (Hashimoto et al. 2012).

This paper's contributions are twofold; **(1)** it proposes a new event causality hypothesis generation method, and **(2)** it shows its effectiveness in a series of large scale experiments.

Even though our target language is Japanese, we believe that our method is applicable to many other languages since most of our ideas are language-independent, as described in Section 2. Examples are translated to English for ease of explanation.

Before ending this section, we clarify our position on causality in this study. As stated above, our definition of causality is based on the probability-raising view of causality against which some criticism exists. Pearl (2000) argued that it is impossible to express that causes raise the probability of their effects within the probability theory framework in the first place. Nevertheless, our definition is viable for the following reason. First, we are generating textual causality knowledge that coincides with the commonsense of ordinary people to develop intelligent natural language processing applications that can reason or infer just like ordinary people. The resulting knowledge base might be inconsistent with so-called scientific truth or the real world itself, but it should help identify the expectations, fears, and interests of ordinary people. For example, we regard causality knowledge *deforestation continues→global warming worsens* as valid since it repeatedly appears in the form of event causality in many web documents. On the other hand, disputes continue about the cause of global warming among experts, and the effect of deforestation on it might actually be deemed negligible in the future. Still, we believe that our system needs causality knowledge to understand the objections against deforestation by ordinary people, for example. Our question is how the causality knowledge (or the beliefs) of ordinary people is expressed in corpora rather than what kind of causality knowledge is scientifically valid. Therefore, the definition of causality using a mathematical theory is secondary. We just used the probability-raising view of causality as criteria for evaluating our results and did not use it in our algorithm in any sense, and indeed, we believe that it worked well as evaluation criteria in our study.

2 Proposed Method

This section details our method. Given event causalities extracted from the web, it generates hypothesis candidates using our semantic relation database in Section 2.1 and ranks them by our hypothesis classifier in Section 2.2. Section 2.3 describes how we prepare the event causalities from which we generate hypotheses.

2.1 Hypothesis Candidate Generation

This section describes our hypothesis candidate generation, which replaces the noun pair (the cause and effect nouns) of an input event causality with another noun pair of the same semantic relation (Figure 1). First, we prepare a **semantic relation database** that records which binary pattern, e.g., *A* CAUSES *B*, which indicates a semantic relation, co-occurs with which noun pairs in 600 million web pages (Akamine et al. 2010).[1] We prepared seven types of binary patterns based on previous work (Hashimoto et al. 2014), as described below. The number in parentheses indicates the number of patterns.

CAUSATION (747) is the causal relation between two entities (e.g., *deforestation* and *global warming*), which is typically expressed by binary pattern *A* CAUSES *B*.

MATERIAL (183) relation holds between a material and a product (e.g., *plutonium* and *atomic bomb*), which can be expressed by *B* IS MADE OF *A*.

NECESSITY (250) can be expressed by *B* REQUIRES *A* with such instances as *verbal aptitude* and *ability to think*.

USE (2,170) holds between the means and the purpose for using something. *A* IS USED FOR *B* is one pattern that can be filled with *e-mailer* and *exchanges of e-mail messages*.

PREVENTION (489) relation can be expressed by *A* PREVENTS *B* with instances like *a mosquito net* and *malaria*.

EXCITATION (55,858) relation is expressed by binary patterns made of excitatory and inhibitory templates (Hashimoto et al. 2012). A template is a predicate with an argument slot like *deploy X* and *avoid X*. Excitatory templates like *deploy X* entail that the function, effect, purpose or role of their argument's referent is activated, enhanced, or manifested, while inhibitory templates like *avoid X* entail that it is deactivated or suppressed. For example, binary pattern *A* LOWERS *B* is made of inhibitory template *lower X*. The excitation relation roughly means that *A* activates (excitatory) or suppresses (inhibitory) *B*. We acquired 43,697 excitatory and inhibitory templates by Hashimoto et al. (2012)'s method[2] and manual annotation for their method's results. Excitatory and inhibitory templates have been successfully applied to various semantic tasks (Oh et al. 2013; Varga et al. 2013; Kloetzer et al. 2013; Tanaka et al. 2013; Sano et al. 2014).

ENTAILMENT (335,780) relation is expressed by binary patterns that have the entailment relation of both directions with one of the binary patterns of the above relations. The aim to incorporate the ENTAILMENT relation is to get broader coverage of noun pairs that are useful for hypothesis generation. Entailment relation binary patterns were collected by a previous method (Kloetzer et al. 2013).

We manually prepared patterns for the first five relations and semi-automatically prepared them for the rest.

[1] These pages were crawled in 2007. We put no restriction on them in terms of domain, discourse type, or writing style.

[2] Hashimoto et al. (2012) constructed a network of templates based on their co-occurrence in web sentences and gave their network a small number of seed templates with the polarity (excitatory or inhibitory) information. Then they inferred the polarity of all the templates in the network using a constraint solver that is based on the spin model (Takamura, Inui, and Okumura 2005).

When checking the co-occurrence of the patterns and the noun pairs that fill in the *A* and *B* slots of the patterns, we consider the dependency structure of sentences in the web pages and the binary patterns using a dependency parser called J.DepP (Yoshinaga and Kitsuregawa 2009). We ignore nouns in our stop-word list that basically consists of word fragments or words that are semantically too vague.

After preparing the semantic relation database, we generate **hypothesis candidates** by replacing the original noun pair of a source event causality with other noun pairs (Figure 1). These other noun pairs must co-occur with the same binary pattern with which the original noun pair co-occurs in our semantic relation database. This process can generate more than one hypothesis from a source event causality, since many noun pairs can co-occur with the same binary pattern. We filter out hypothesis candidates (phrase pairs) if they consist of a phrase whose noun and template have fewer than ten dependency relations in the web pages. For example, for hypothesis candidate *deploy a security camera→avoid crime*, we check the occurrence frequency of a dependency relation between *deploy X* and *a security camera* and between *avoid X* and *crime*.

Then we apply the following three filters to the generated hypothesis candidates for better precision: (1) the PMI filter keeps only hypotheses whose noun pair (cause and effect nouns) is registered in the word co-occurrence frequency database[3] released by the ALAGIN forum[4] and its PMI value is greater than or equal to zero. The database records 5,000 words of the largest PMI values for each entry word. The number of entry words is about 500,000, and the PMI values were calculated using about 100 million Japanese web pages with a co-occurrence window set to four sentences. (2) The stop-word filter keeps only hypotheses whose cause and effect nouns are not in our stop-word list. (3) The same-noun filter keeps only hypotheses whose cause and effect nouns are different, since we target hypotheses that describe causal relations between two different entities.

Finally, we keep only hypotheses that do not exist in source event causalities (phrase pairs). For example, if we generate hypothesis *deploy a mosquito net→avoid malaria* that is included in the source event causalities, it is discarded. We call the remaining hypotheses *phrase pair level novelty hypotheses*. Optionally, we further keep only hypotheses whose noun pairs (cause and effect nouns) do not exist in the noun pair list from the source event causalities, and discard the others. For example, if we generate the above *mosquito net* hypothesis but one of source event causalities consists of noun pair (*a mosquito net* and *malaria*), e.g., *use a mosquito net→prevent malaria*, the hypothesis is discarded. We call these remaining hypotheses *noun pair level novelty hypotheses*.

One might question the value of generating causality hypothesis phrase pairs through binary patterns that already indicate causality to some extent. Indeed, such binary patterns as *A* CAUSES *B* express causality, e.g., *earthquakes cause tsunamis.* However, although there are many events involving earthquakes like *worry about earthquakes, predict earthquakes,* and *trigger earthquakes,* binary pattern-based causalities like *earthquakes cause tsunamis* cannot indicate which events involving earthquakes cause tsunamis. With our proposed method, we can identify such events by properly translating binary pattern-based causalities into causality phrase pairs: *trigger earthquakes→cause tsunamis.* This is important for textual causal inference. If an inference system knows that someone is worried about earthquakes and only has causality knowledge *earthquakes cause tsunamis,* it might wrongly infer that a tsunami will be caused. If the system can hypothesize *trigger earthquakes→cause tsunamis* from *earthquakes cause tsunamis,* it will not make such a mistake. Thus, generating causality hypothesis phrase pairs through binary pattern-based causalities by augmenting predicates is critical from an application's perspective.

2.2 Hypothesis Ranking

Some hypothesis candidates are more plausible for event causality than others. We identify plausible hypothesis candidates by an SVM classifier (HYPOCLASSIFIER, hereafter), which is trained by labeled data and features that Hashimoto et al. (2014) developed for event causality extraction tasks. In this section, we describe HYPOCLASSIFIER, which takes a holistic approach to identifying plausible event causality hypotheses. It checks the semantic relations between two nouns (cause and effect nouns), the semantic class of each noun, predicate semantics, and association strength between words, among others. The following are its **features**: (for more details, see Hashimoto et al. (2014)): Semantic relation features embody an assumption that if two nouns in a causality (hypothesis) candidate (e.g., *slash-and-burn agriculture* and *desertification* in *conduct slash-and-burn agriculture→exacerbate desertification*) take a specific semantic relation (e.g., *A* CAUSES *B*), the candidate tends to be plausible. As semantic relations, we use those previously described (Hashimoto et al. 2014).

Context features represent the likely contexts in which event causality might appear, including causal connectives, distances between elements of event causalities, and words in context. These features are extracted from the original sentences from which the event causalities were extracted. These original sentences include not only phrase pairs that represent event causality but also connectives between the two phrases. Notice that our causality hypotheses do not have the original sentences in which they are written since they are generated rather than extracted from the corpora, making utilization of the context features impossible. To use HYPOCLASSIFIER with minimum costs, we automatically create *artificial* original sentences for the generated hypotheses. These artificial original sentences contain only two phrases of a hypothesis that are placed side by side (the cause phrase comes first) with a connective (which roughly corresponds to "*and*") between them. For example, for *deploy a security camera→avoid crime*, we create the following artificial original sentence: *we deploy a security camera and avoid crime*.

[3]https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-list.html, ID: A-5

[4]http://alagin.jp/index-e.html

Association features are based on an assumption that each word of the cause phrase must have a strong association (i.e., PMI) with the effect phrase.

Base features include nouns, their semantic classes (Kazama and Torisawa 2008), templates, and their excitation polarities (Section 2.1), among others.

Labeled data consist of 147,519 examples (15,195 are positive). Using the features and the labeled data, HYPOCLASSIFIER is trained by SVM-Light with polynomial kernel $d = 2$ (svmlight.joachims.org).

Ranking is based on SVM scores (the distance from the SVM hyperplane), which represent the plausibility of the hypothesis candidates as event causalities.

2.3 Event Causality Extraction

To extract the source event causalities from which we generate hypotheses, we follow Hashimoto et al. (2014), who extracted phrase pairs as event causality candidates from single sentences in 600 million web pages. Each phrase of the phrase pairs must consist of an excitatory or inhibitory template (Section 2.1) and a noun that fills its slot. The predicate of the cause phrase must syntactically depend on that of the effect phrase, and the cause phrase must precede the effect phrase in a sentence, since the temporal order between events is usually determined by precedence in a Japanese sentence.

In this way, we extracted 132,528,706 event causality candidates. To them, following Hashimoto et al. (2014), we applied filters, including the three described in Section 2.1 and those checking the context of such event causality candidates as the connective between the cause and effect phrases. 2,451,254 (2.4M) event causalities remained to which we applied Hashimoto et al. (2014)'s event causality classifier to rank them by the distance from the SVM hyperplane. We report the precision of this method in Section 3.2.

3 Experiments

Through a series of large-scale experiments with 70% precision, our method generated **(i)** 347,093 noun pair level novelty hypotheses and **(ii)** 302,350 phrase pair level novelty hypotheses from the 2.4M source event causalities (for these novelty criteria, see Section 2.1). **(iii)** Our semantic relations are actually useful for hypothesizing event causalities. **(iv)** Our method outperforms a state-of-the-art hypothesis generation method. **(v)** Regarding event causality *acquisition* (i.e., either extraction or generation), our method outperformed the state-of-the-art event causality extraction method (Hashimoto et al. 2014).

First, we evaluate our method in the noun pair level novelty setting and support claims (i) and (iii) (Section 3.1). In the noun pair level novelty setting, the generated hypotheses are removed (i) if their phrase pairs are found in the 132,528,706 event causality candidates (Section 2.3) or (ii) if their noun pairs exist in the range of 60% precision of all of the 2.4M source event causalities or in Hashimoto et al. (2014)'s labeled data.

Next we evaluated our method in the phrase pair level novelty setting, which supports claims (ii), (iv), and (v) (Section 3.2). In the phrase pair level novelty setting, we removed the generated hypotheses if their phrase pairs exist in the 132,528,706 event causality candidates (Section 2.3).

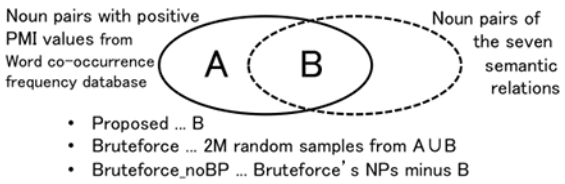

Figure 2: Relationship of the three methods.

Three human annotators (not the authors) did the evaluations and determined the label of each hypothesis by majority vote. The kappa (Fleiss 1971) of their judgments was 0.55, which is moderate agreement (Landis and Koch 1977). As with Hashimoto et al. (2014), our event causality hypotheses must be *self-contained*, i.e., intelligible as event causalities by themselves without contextual information, since one of our main objectives is future scenario generation (Hashimoto et al. 2014) for which event causalities and hypotheses must be self-contained. For example, the *mosquito net* example in Section 1 is self-contained, but *cause disease→develop hypertension* is not since it is unclear whether the *disease* is relevant to blood pressure by itself, and we cannot judge its plausibility.

3.1 Noun Pair Level Novelty Setting

We compare the following five **methods**. Further details for each method are given below.

`Proposed` is our method.

`Bruteforce` is identical to `Proposed` except that it does not consider the semantic relations between nouns and basically uses any noun pairs of any semantic relations. It uses two million noun pairs randomly sampled from all the noun pairs with positive PMI values in the word co-occurrence frequency database (Section 2.1).

`Bruteforce`$_{noBP}$ is identical to `Bruteforce` except that it does not use noun pairs that co-occur with one of the binary patterns in our semantic relation database.

`Random`$_{Proposed}$ is identical to `Proposed` except that it does not rank the generated hypotheses.

`Random`$_{Bruteforce}$ is identical to `Bruteforce` except that it does not rank the generated hypotheses.

Figure 2 untangles the relationship of three methods: `Proposed`, `Bruteforce`, and `Bruteforce`$_{noBP}$. The oval with solid lines (A∪B) represents the noun pairs with positive PMI values, which are used for the PMI filter (Section 2.1). The oval with dashed lines represents the noun pairs that co-occur with our binary patterns of the seven relation types. `Proposed` generated hypotheses using noun pairs in B. `Bruteforce`'s noun pairs are the two million random samples from A∪B, and `Bruteforce`$_{noBP}$'s noun pairs are `Bruteforce`'s noun pairs minus B.

Our **intentions** underlying this experimental design is to confirm the following. **(a)** If `Proposed` outperforms

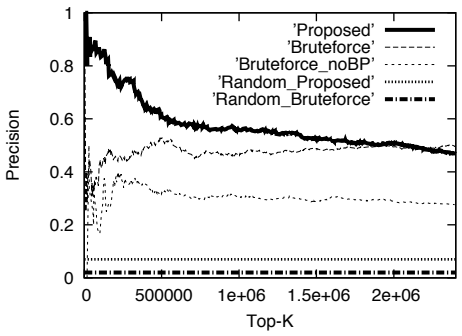

Figure 3: Results in noun pair level novelty setting.

Bruteforce, it indicates that it is effective to use *only* noun pairs bearing the seven types of semantic relations in this paper. **(b)** If Bruteforce outperforms $\texttt{Bruteforce}_{noBP}$, Bruteforce's (reasonably good) performance may well be attributed to its noun pairs that happen to bear our semantic relations. Perhaps HYPOCLASSIFIER ranks these hypotheses higher with noun pairs bearing our semantic relations, since it considers the semantic relations between nouns. If Bruteforce outperforms $\texttt{Bruteforce}_{noBP}$, that result indicates that it is important to consider the semantic relations between nouns to rank the generated hypotheses. **(c)** If $\texttt{Random}_{Proposed}$ and $\texttt{Random}_{Bruteforce}$ work poorly, it indicates that it is indispensable to rank the hypotheses. If all are confirmed, we can conclude that our semantic relations are useful for hypothesis generation.

The results (precision for the top ranked hypothesis candidates) are shown in Figure 3. Proposed outperformed the others and generated 347,093 noun pair level novelty hypotheses with 70% precision from the 2.4M source event causalities (Claim **(i)**). Since Bruteforce outperformed $\texttt{Bruteforce}_{noBP}$ and the two random methods worked very poorly, we conclude that our semantic relations are actually useful for hypothesis generation (Claim **(iii)**).

Below are further **details** for each method.

Proposed generated 83,468,106 hypotheses, and we evaluated 500 random samples from the top 2.4 million hypotheses. Among the 2.4 million, there were 127,318 different noun pairs. Examples of Proposed's hypotheses are given below. For each example, a hypothesis, its source event causality, a binary pattern through which the hypothesis was generated from the source, and the semantic relation type of the binary pattern are shown in this order. '*A*' and '*B*' represent the cause and effect nouns respectively. The number indicates the rank of the hypothesis.

105: *increase in elastin→improve elasticity* was generated from *increase in dopamine→improve happiness* through *A* PROVIDING *B* (ENTAILMENT relation: *A* PROVIDING *B* has an entailment relation of both directions with EXCITATION relation pattern *A* BRINGS ABOUT *B*).

263: *cause static electricity→cause malfunction* is generated from *cause heavy rain→cause flood damage* through *B* BY *A* (CAUSATION relation).

30,588: *monetary relaxation continues→leads to a drop in the yen* is generated from *declining birth rate continues→leads to abolishment of schools* through *A* FACILITATES *B* (EXCITATION relation).

Bruteforce generated 87,974,520 hypotheses. We prepared noun pairs for it as follows. First, we obtained 2,018,170,662 noun pairs with positive PMI values from the word co-occurrence frequency database (Section 2.1). Then we randomly sampled two million of them. We evaluated 500 random samples from the top 2,400 hypotheses, since we only used two million noun pairs out of 2,018,170,662 (about $\frac{1}{1,000}$) and thus the top 2,400 hypotheses correspond to the top 2.4 million if we used all the 2,018,170,662 noun pairs. In the top 2,400 hypotheses, there were 364 different noun pairs, and thus we estimate that there are 364,000 ($364 \times 1,000$) different noun pairs in its top 2.4 million hypotheses. Bruteforce emulates a brute force search over the possible event causality hypothesis space, although we restricted noun pairs to those with positive PMI values and due to the limitation of machine resources and time constraints, only used randomly sampled two million noun pairs. Despite its simplicity, this method showed a relatively good performance. Proposed, on the other hand, efficiently reduced the search space by semantic relations and indeed outperformed Bruteforce for around the top two million pairs.

$\texttt{Bruteforce}_{noBP}$ generated 84,802,130 hypotheses, and we evaluated 500 random samples from its top 2,400 hypotheses, where there were 118 different noun pairs.

3.2 Phrase Pair Level Novelty Setting

We compared the following three **methods**. Further details are given below.

Proposed is our method.

CHG is the state-of-the-art event causality hypothesis generation method (Hashimoto et al. 2012), which we call Contradiction-based Hypothesis Generation (CHG). It generates a hypothesis by replacing each phrase of a source event causality with another phrase that semantically contradicts with the original phrase.

SrcEC is the 2.4M source event causalities ranked by the SVM scores of Hashimoto et al.'s event causality classifier.

Our **intention** behind this experimental setting is to confirm that Proposed outperforms CHG, the state-of-the-art event causality hypothesis generation method and that even in event causality *acquisition* tasks (either extraction or generation), Proposed outperforms SrcEC, the state-of-the-art event causality extraction method. Note that we did not filter out any of SrcEC's output, although its output is event causalities extracted from the web and hence is not a novelty hypothesis. If these are confirmed, we can conclude that our

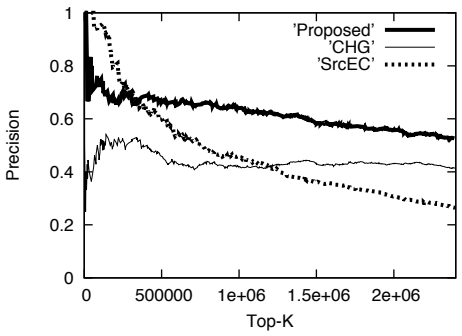

Figure 4: Results in phrase pair level novelty setting.

claims **(iv)** and **(v)** (in the first paragraph of Section 3) are valid.

The results are shown in Figure 4. `Proposed` generated 302,350 phrase pair level novelty hypotheses from the source event causalities with 70% precision (Claim **(ii)**) and outperformed the other two methods (Claims **(iv)** and **(v)**).

Below are further **details** for each method.

`Proposed` generated 170,584,994 hypotheses. This number is different from that in Section 3.1 because the novelty criterion is different. We evaluated 500 random samples from the top 2.4 million hypotheses, which consisted of 75,249 different noun pairs. Below are examples of `Proposed`'s hypotheses. The format is the same as the examples in Section 3.1.

252: *stomach acid increases→develops into a (gastric) ulcer* is generated from *neutral fat increases→develops into hyperlipemia* through *A* WHICH IS A CAUSE OF *B* (CAUSATION relation).

933: *use collagen→produce beautiful skin* is generated from *use an accelerator→produce neutrinos* through *B* ARE MADE BY *A* (MATERIAL relation).

10,287: *cause stroke→lead to speech difficulty* is generated from *cause rising air→lead to cumulonimbus clouds* through *B* OCCURS BY *A* (EXCITATION relation).

841,893 *increase in nuclear power plants→radioactive pollution occurs* is generated from *increase in plankton population→red tide occurs* through *A* CAUSING *B* (CAUSATION relation).

`CHG` assumes that if a source event causality is valid, its inverse is often valid as well. Note that a hypothesis generated by `CHG` has the same noun pair as the source event causality. For example, given event causality *get periodontal disease→have bad breath*, `CHG` generates hypothesis *cure periodontal disease→prevent bad breath*, where phrase pairs *get periodontal disease* $\perp$ *cure periodontal disease* and *have bad breath* $\perp$ *prevent bad breath* are both contradictory ($\perp$ indicates that two phrases contradict each other). To acquire contradiction phrase pairs, we followed Hashimoto et al. (2012), who extracted two phrases as a contradiction pair if (a) their templates had opposite excitatory and inhibitory polarities, e.g., *have X* (excitatory) and *prevent X* (inhibitory) (Section 2.1), (b) they shared the same argument noun, e.g., *bad breath*, and (c) the part-of-speech of their predicates was the same. Then a phrase pair (p, q) was given as a contradiction score:

$$Ct(p, q) = |s_p| \times |s_q| \times sim(t_p, t_q).$$

Here, t_p and t_q are the templates of p and q, $|s_p|$ and $|s_q|$ are the absolute scores of t_p and t_q's excitation values, whose range is $[-1, 1]$ and are positive if the template is excitatory and negative if it is inhibitory (see Hashimoto et al. (2012) for the definition of excitation value), and $sim(t_p, t_q)$ is a distributional similarity score, which was calculated by Hashimoto et al. (2009) in our study. Based on these contradiction phrase pairs, hypothesis $q_1 \rightarrow q_2$ generated from $p_1 \rightarrow p_2$ is given by the following hypothesis score:

$$Hp(q_1, q_2) = Ct(p_1, q_1) \times Ct(p_2, q_2) \times Cs(p_1, p_2).$$

Here, $p_1 \perp q_1$ and $p_2 \perp q_2$ are contradiction pairs, $Ct(p_1, q_1)$ and $Ct(p_2, q_2)$ are their contradiction scores, and $Cs(p_1, p_2)$ is the event causality score for $p_1 \rightarrow p_2$. Regarding excitatory and inhibitory templates which are necessary for `CHG`, we first obtained 8,686 templates with excitation values from the 600 million web pages based on Hashimoto et al. (2012). Then we augmented these templates with several types of negation form versions and obtained 60,756 templates. For a fair comparison with `Proposed`, we used the 2.4M event causalities as the source of the hypotheses. Also, we filtered out the hypothesis candidates (phrase pairs) if they consisted of a phrase whose noun and template had fewer than ten dependency relations in the web pages, in the same way as in `Proposed`. `CHG` does not need the filters that `Proposed` applies to its hypothesis candidates, since all of its noun pairs are the same as those of the source event causalities; such noun pairs have already survived the filters. See Section 2.3. 11,013,360 hypotheses were generated. We evaluated 500 random samples from the top 2.4 million hypotheses, in which there were 27,571 different noun pairs.

`SrcEC` is the 2.4M (2,451,254, to be precise) source event causalities, ranked by the SVM scores of Hashimoto et al.'s event causality classifier. We evaluated 500 random samples from the top 2.4 million event causalities, in which there were 1,856,836 different noun pairs. `Proposed` and `CHG` outperformed `SrcEC` except for the top 150,000 or so. This is because the two hypothesis generation methods tended to pick up a relatively small number of noun pairs that seemed to lead to plausible hypotheses and generated many hypotheses from these noun pairs with different template pairs. For example, from noun pair *global warming* and *abnormal climate*, `Proposed` generated hypotheses *global warming increases→lead to abnormal climate*, *global warming occurs→cause abnormal climate*, and *global warming worsens→abnormal climate continues*. In other words, these hypothesis generation methods tend to provide many *paraphrases* of causality hypotheses, which we believe are beneficial to intelligent natural language processing tasks (Iordanskaja, Kittredge, and

Polguère 1991; Lin and Pantel 2001; McKeown et al. 2002; Ravichandran and Hovy 2002; Kauchak and Barzilay 2006; Callison-Burch, Koehn, and Osborne 2006).

3.3 Discussion

In this section we present error analyses for our proposed method and discuss its limitations. From the experiment results in Section 3.1, we noticed that its errors were due mainly to errors in the preprocessing stages, such as **(a)** semantic relation database preparation (Section 2.1) and **(b)** source event causality extraction (Section 2.3).

As for the above (a), take erroneous (or nonsense) event causality hypothesis *unequal settling worsens→ground becomes severe* as an example, which was generated from valid event causality *alveolar pyorrhea worsens→bad breath becomes severe* through semantic relation *A* IS THE CAUSE OF *B*, where *unequal settling* and *alveolar pyorrhea* correspond to *A* and *ground* and *bad breath* correspond to *B*. Notice that our semantic relation database preparation method wrongly extracted the triple (*A* IS THE CAUSE *B*, A=*unequal settling*, B=*ground*) from our corpus, which should more accurately be something like: *A* IS THE CAUSE *B*, A=*unequal settling*, B=*the ground's inclination*. Other cases include the errors introduced in the entailment relation binary pattern acquisition (Section 2.1).

As for the above (b), take erroneous event causality hypothesis *neutralize impure substances→have off-flavor* as an example, which was generated from erroneous event causality *neutralize gastric acid→have gastric ulcer* through semantic relation *A* IS A SOURCE OF *B*, where *impure substances* and *gastric acid* correspond to *A* and *off-flavor* and *gastric ulcer* correspond to *B*. Many of the above preprocessing stages' errors were due to the dependency parser error.

An obvious limitation of our method is that it cannot generate event causality hypotheses in which more than two events (or more than two nouns) are involved, since our method only considers two nouns of cause and effect phrases. In this world, however, there are event causalities in which more than one cause events must occur for a corresponding effect event to occur. Our future work includes the extension of our proposed framework to deal with such event causalities.

4 Related Work

Many methods have been proposed to *extract* event causalities from corpora (Abe, Inui, and Matsumoto 2008; Bethard and Martin 2008; Radinsky, Davidovich, and Markovitch 2012; Oh et al. 2013; Torisawa 2006; Riaz and Girju 2010; Do, Chan, and Roth 2011; Torisawa 2006; Chambers and Jurafsky 2008; 2009; Hashimoto et al. 2012; 2014). The problem, however, is that it is unlikely that all the event causalities that we recognize in this world are written in corpora. Therefore, we need a method that acquires event causality knowledge that is not written in corpora. Our proposed method *synthesizes* a large number of pieces of knowledge like semantic relations and event causalities that are scattered across the web to hypothesize plausible event causalities. It conducts a generate-and-test search for plausible event causality hypotheses by exploiting semantic relations to effectively reduce the search space.

Actually, some methods (have the potential to) acquire event causalities that are not written in corpora. Hashimoto et al. (2012) proposed a hypothesis generation method (`CHG`) that exploits the contradiction knowledge between phrases. Our method outperformed `CHG` by a large margin. One interesting research direction is to integrate Hashimoto et al. (2012)'s idea of exploiting the contradiction knowledge in our method.

Radinsky et al. (2012) target news domains and induce event causality rules like *if an earthquake occurs next to an island, a tsunami warning will be issued for its nearest ocean* from event causalities extracted from past newspapers. They generate hypotheses (predicting news events) using the rules, but their hypotheses are inevitably "chained" to the rules; they can only generate hypotheses that are (quite) similar to previously observed causalities. On the other hand, our method can "leap" from source event causalities due to our semantic relation-based framework and generate very different hypotheses from the source event causalities, as in the above examples. Our method is also domain-independent.

Tanaka et al. (2012) induce event causality rules from event causalities extracted by exploiting deverbal nouns. Even though they have not reported it, they could have generated hypotheses by the rules. However, they do not give any ranking scheme for their generated hypotheses, which is indispensable, as we showed in Section 3.1; methods without proper ranking, $\texttt{Random}_{Proposed}$ and $\texttt{Random}_{Bruteforce}$, performed very poorly.

Carlson et al. (2010) automatically augment existing knowledge bases of the semantic categories of nouns and the semantic relations between nouns, which shares a similar spirit with our study.

5 Conclusion

We proposed a method of hypothesizing plausible event causality hypotheses from event causalities extracted from the web by exploiting semantic relations. With 70% precision, our method generated 347,093 noun pair level novelty hypotheses and 302,350 phrase pair level novelty hypotheses from the 2.4M event causalities extracted from the web. Our method outperformed the state-of-the-art hypothesis generation method by a large margin.

We are planning to release generated event causality hypotheses to the public in the near future.

References

Abe, S.; Inui, K.; and Matsumoto, Y. 2008. Two-phrased event relation acquisition: Coupling the relation-oriented and argument-oriented approaches. In *COLING 2008*, 1–8.

Akamine, S.; Kawahara, D.; Kato, Y.; Nakagawa, T.; Leon-Suematsu, Y. I.; Kawada, T.; Inui, K.; Kurohashi, S.; and Kidawara, Y. 2010. Organizing information on the web to support user judgments on information credibility. In *IUCS 2010*, 122–129.

Bethard, S., and Martin, J. H. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. In *ACL-08: HLT (Short paper)*, 177–180.

Callison-Burch, C.; Koehn, P.; and Osborne, M. 2006. Improved statistical machine translation using paraphrases. In *HLT-NAACL 2006*, 17–24.

Carlson, A.; Betteridge, J.; Kisiel, B.; Settles, B.; Jr., E. R. H.; and Mitchell, T. M. 2010. Toward an architecture for never-ending language learning. In *AAAI 2010*.

Chambers, N., and Jurafsky, D. 2008. Unsupervised learning of narrative event chains. In *ACL-08: HLT*, 789–797.

Chambers, N., and Jurafsky, D. 2009. Unsupervised learning of narrative schemas and their participants. In *ACL-IJCNLP 2009*, 602–610.

Do, Q. X.; Chan, Y. S.; and Roth, D. 2011. Minimally supervised event causality identification. In *EMNLP 2011*, 294–303.

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

Hashimoto, C.; Torisawa, K.; Kuroda, K.; Murata, M.; and Kazama, J. 2009. Large-scale verb entailment acquisition from the web. In *EMNLP 2009*, 1172–1181.

Hashimoto, C.; Torisawa, K.; Saeger, S. D.; Oh, J.-H.; and Kazama, J. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *EMNLP-CoNLL 2012*, 619–630.

Hashimoto, C.; Torisawa, K.; Kloetzer, J.; Sano, M.; Varga, I.; Oh, J.-H.; and Kidawara, Y. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *ACL 2014*, 987–997.

Iordanskaja, L.; Kittredge, R.; and Polguère, A. 1991. Lexical selection and paraphrase in a meaning-text generation model. In Paris, C. L.; Swartout, W. R.; and Mann, W. C., eds., *Natural language generation in artificial intelligence and computational linguistics*. Kluwer Academic Press. 293–312.

Kauchak, D., and Barzilay, R. 2006. Paraphrasing for automatic evaluation. In *HLT-NAACL 2006*, 455–462.

Kazama, J., and Torisawa, K. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *ACL-08: HLT*, 407–415.

Kloetzer, J.; Saeger, S. D.; Torisawa, K.; Hashimoto, C.; Oh, J.-H.; and Ohtake, K. 2013. Two-stage method for large-scale acquisition of contradiction pattern pairs using entailment. In *EMNLP 2013*, 693–703.

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

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

McKeown, K. R.; Barzilay, R.; Evans, D.; Hatzivassiloglou, V.; Klavans, J. L.; Nenkova, A.; Sable, C.; Schiffman, B.; and Sigelman, S. 2002. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the 2nd international conference on Human Language Technology Research*, 280–285.

Oh, J.-H.; Torisawa, K.; Hashimoto, C.; Sano, M.; Saeger, S. D.; and Ohtake, K. 2013. Why-question answering using intra- and inter-sentential causal relations. In *ACL 2013*, 1733–1743.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Radinsky, K.; Davidovich, S.; and Markovitch, S. 2012. Learning causality for news events prediction. In *WWW 2012*, 909–918.

Ravichandran, D., and Hovy, E. H. 2002. Learning surface text patterns for a question answering system. In *ACL 2002*, 41–47.

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

Sano, M.; Torisawa, K.; Kloetzer, J.; Hashimoto, C.; Varga, I.; and Oh, J.-H. 2014. Million-scale derivation of semantic relations from a manually constructed predicate taxonomy. In *COLING 2014*, 1423–1434.

Takamura, H.; Inui, T.; and Okumura, M. 2005. Extracting semantic orientation of words using spin model. In *ACL 2005*, 133–140.

Tanaka, M.; De Saeger, S.; Ohtake, K.; Hashimoto, C.; Hijiya, M.; Fujii, H.; and Torisawa, K. 2013. WISDOM2013: A large-scale web information analysis system. In *IJCNLP 2013 (Demo Track)*, 45–48.

Tanaka, S.; Okazaki, N.; and Ishizuka, M. 2012. Acquiring and generalizing causal inference rules from deverbal noun constructions. In *COLING 2012*, 1209–1218.

Torisawa, K. 2006. Acquiring inference rules with temporal constraints by using japanese coordinated sentences and noun-verb co-occurrences. In *HLT-NAACL 2006*, 57–64.

Varga, I.; Sano, M.; Torisawa, K.; Hashimoto, C.; Ohtake, K.; Kawai, T.; Oh, J.-H.; and Saeger, S. D. 2013. Aid is out there: Looking for help from tweets during a large scale disaster. In *ACL 2013*, 1619–1629.

Yoshinaga, N., and Kitsuregawa, M. 2009. Polynomial to linear: Efficient classification with conjunctive features. In *EMNLP 2009*, 542–1551.