Toward Finding Semantic Relations not Written in a Single Sentence: An Inference Method using Auto-Discovered Rules

Masaaki Tsuchida†§* **Kentaro Torisawa**‡ **Stijn De Saeger** ‡
Jong-Hoon Oh‡ **Jun'ichi Kazama**‡ **Chikara Hashimoto**‡ **Hayato Ohwada**§
† Information and Media Processing Laboratories, NEC Corporation, Nara, Japan
m-tsuchida@cq.jp.nec.com
‡ Information Analysis Laboratory, National Institute of
Information and Communications Technology, Kyoto, Japan
{torisawa,stijn,rovellia,kazama,ch}@nict.go.jp
§ Graduate school of Science and Technology, Tokyo University of Science, Chiba, Japan.
ohwada@ia.noda.tus.ac.jp

Abstract

Recent advances in automatic knowledge acquisition methods have enabled us to construct massive knowledge bases of semantic relations. Most previous work has focused on semantic relations explicitly expressed in single sentences. Our goal in this work is to obtain valid *non-single sentence* relation instances, which are not written in any single sentence and may not be even written in a large corpus. We develop a method to infer new semantic relation instances by applying auto-discovered inference rules, and show that our method inferred a considerable number of valid instances that were not written in single sentences even in 600 million Web pages.

1 Introduction

Recent advances in automatic relation acquisition methods (Agichtein and Luis, 2001; Etzioni et al., 2004; Pantel and Pennacchiotti, 2006; Paşca et al., 2006; Banko and Etzioni, 2008; De Saeger et al., 2009) have opened the way to build large knowledge bases containing a huge number of semantic relation instances such as "CAUSE(*allergen*, *allergy*)" and "PREVENTION(*coffee*, *drowsiness*)". Such a massive knowledge base is valuable for applications like innovation and risk management (Torisawa et al., 2010), like finding potential and unexpected causes of a disease, unknown side-effects of a drug, and so on. In such tasks overlooking a small piece of information can have grave consequences.

In this work, our goal is to acquire *Non-single Sentence relation instances* (**NS instances**), which are not written in any single sentences, in order to reduce the possibility of overlooking valuable information. More precisely, if the two nouns in an instance (e.g., "allergen" and "allergy" in "CAUSE(*allergen*, *allergy*)") do not appear in any single sentences in a given corpus together with the other clues indicating the relation type (e.g., the verb "cause"), the instance is an NS instance. *NS instances* may be *indirectly* written in a sequence of several sentences using anaphora or may not be written at all even in a given corpus. In the following, we call the complement of NS instances *Single Sentence relation instances* (**SS instances**), which consist of two nouns that co-occur in some single sentences in a given corpus together with some evidence for a relation type.

Most existing relation acquisition methods acquire relation instances using lexico-syntactic patterns (Pantel and Pennacchiotti, 2006; Paşca et al., 2006; De Saeger et al., 2009) such as "X causes Y" or probabilistic sequence labeling models (Banko and Etzioni, 2008). These methods basically rely on the structure of single sentences and they are for acquiring SS instances. Thus we consider that NS instances are practically beyond their reach. A few attempts to overcome this limitation include inference-based methods. These methods take SS instances provided by other methods as input and infer relation instances including NS ones using auto-discovered inference rules (Schoenmackers et al., 2010; Carlson et al., 2010) or distributional similarities (Tsuchida et al., 2010). We consider such inference approaches to be promising for acquiring NS instances.

This paper proposes an inference-based method for acquiring NS instances. We start from a set of *seed relation instances* of a target relation, which are acquired by an existing semi-supervised

*This work was done when the author was at the National Institute of Information and Communications Technology.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 902–910,
Chiang Mai, Thailand, November 8 – 13, 2011.

pattern-based method (De Saeger et al., 2009), and induce a large number of *recursive* inference rules to infer new relation instances. The instances inferred by the rules are ranked based on the scores assigned to the rules and the top instances are produced as final output.

Specifically, our method infers new instances of particular *target* relations by applying the following type of *recursive rules* to large corpora.

$$\text{“X } \textit{pattern} \text{ Y”} \wedge \mathrm{R}_{\mathrm{SEED}}(\mathrm{Y,Z}) \rightarrow \mathrm{R}_{\mathrm{HYPO}}(\mathrm{X,Z})$$

Here, “X *pattern* Y” indicates that X and Y co-occur in a certain lexico-syntactic pattern, $\mathrm{R}_{\mathrm{SEED}}$ (Y,Z) is one of the seed relation instances, and an inferred instance of the target relation is denoted by $\mathrm{R}_{\mathrm{HYPO}}$ (X,Z). Though we distinguish $\mathrm{R}_{\mathrm{SEED}}$ (Y,Z) and $\mathrm{R}_{\mathrm{HYPO}}$ (X,Z) by the subscripts “SEED” and “HYPO”, they are supposed to be the same target relation, and thus the rules are recursive. Note that this type of recursive rules could not be employed in an existing inference-based method (Schoenmackers et al., 2010) as described in Section 4.1.

For acquiring *NS instances*, rules combine patterns and relation instances (“X *pattern* Y” and $\mathrm{R}_{\mathrm{SEED}}$ (Y,Z)) scattered throughout many distinct documents. The following is an example rule learned by our method.

$$\text{”X } \textit{is rich in} \text{ Y”} \wedge \mathrm{PREVENTION}_{\mathrm{SEED}}(\mathrm{Y,Z}) \rightarrow \mathrm{PREVENTION}_{\mathrm{HYPO}}(\mathrm{X,Z})$$

This can be interpreted as: “If X *is rich in* Y and Y *prevents* Z, then X *prevents* Z.”, and seems valid in many cases. An interesting example is that the rule generated the relation instance “PREVENTION(*X=blue-backed fish*, *Z=cerebral thrombosis*)“ where “*Y=eicosapentaenoic acid*”. This is a rather well-known fact in Japan these days, and you can find many Web pages stating this fact in early 2011. However, the words “*blue-backed fish*” and “*cerebral thrombosis*” do not co-occur in any four sentence window in the 600M Web page corpus used in our experiments, which was crawled in 2008. Thus, “PREVENTION(*blue-backed fish*, *cerebral thrombosis*)“ is an NS instance and exemplifies the importance and necessity of inferring NS instances.

This example also introduces the idea underlying our evaluation scheme. Proper evaluation of new relation hypotheses that are not mentioned together in the extraction corpus would require verification by domain experts, which is clearly beyond our resources. We therefore evaluate new relation hypotheses using a larger and newer corpus (i.e. a commercial search engine), under the naive assumption that *correct* instances will be found in this bigger corpus. We realize this evaluation scheme cannot do justice to genuinely unknown instances but only gives a lowerbound of the true precision.

Through a series of experiments for acquiring three semantic relations, *causality*, *prevention*, and *material*, from a 600 million page Japanese Web corpus, we show that our method can infer *valid* NS instances. We also compare our method to the markov logic inference algorithm introduced in Schoenmackers et al. (2010) and show that our procedure, while considerably simpler, outperforms it.

2 Related Work

Previous work can be categorized into three groups: 1) methods for *extracting* SS instances (Etzioni et al., 2004; Pantel and Pennacchiotti, 2006; Paşca et al., 2006; Banko and Etzioni, 2008; De Saeger et al., 2009), 2) methods for *inferring* relation instances (Carlson et al., 2010; Schoenmackers et al., 2010; Tsuchida et al., 2010), and 3) work in bioinformatics aiming at helping users to discover unseen relation instances as novel knowledge (Swanson, 1986; Srinivasan, 2004; Hu et al., 2006; Hristovski et al., 2008).

The methods in the first category aim at extracting a large number of instances by employing automatically induced *paraphrases* of lexico-syntactic patterns or probabilistic sequence labeling methods. In this category, we focus on De Saeger et al. (2009), which is the seed instance extractor employed in this work. This takes *seed patterns*, such as “X causes Y”, as input and learns a large amount of *paraphrase patterns* to extract relation instances from a corpus. Unlike other pattern-based methods, De Saeger et al. (2009) learns *class dependent paraphrase patterns*, which place semantic word class restrictions on the noun pairs they may extract, like “*X:*`chemical` causes *Y:*`disease`”. These class restrictions enable to distinguish between multiple senses of frequent but highly ambiguous patterns. For instance, given a class *independent* pattern “X causes Y” as seed, if we restrict X and Y in “Y from X” to the classes of chemicals and diseases (as in “*cancer* from *cadmium*”), the class dependent pattern “*Y:*`disease` from *X:*`chemical`” becomes a valid paraphrase of “X causes Y”. Note that, other class restric-

tions of the same pattern (e.g., "*Y:*products from *X:*company", as in "iPhone from Apple") may not yield a valid paraphrase of "X causes Y". To obtain word classes they use a large-scale word clustering algorithm (Kazama and Torisawa, 2008), and rank each instance in the corpus according to a score based on the semantic similarity between the seed patterns and each class dependent pattern the instance co-occurs with.

Although much work in this category successfully extracts the instances *implicitly* written in a single sentence based on a wide range of non-trivial evidence including paraphrases (e.g., "Y by X" for causality), the applicability of these methods is restricted to SS instances.

In the second category, Schoenmackers et al. (2010), which is the most relevant to our work, takes relation instances provided by TextRunner (Banko and Etzioni, 2008) as input and induces Horn clauses as inference rules. The weights of the rules and the probabilities of the hypothesized instances are estimated by a markov logic network (Richardson and Domingo, 2006; Huynh and Mooney, 2008). They also proposed a weighted counting method for discounting the effects of uncertain (or infrequent) instances, and a method for discounting weights of longer rules by strong Gaussian prior.

The goal of Schoenmackers et al. (2010) was to acquire *implicit* relation instances. Note that their "implicit" instances cover what we call SS and NS instances, while our target are only NS instances. For instance, they regarded the instances acquired by simple paraphrase rules such as "CAN_CAUSE(X,Y) $\rightarrow$ CAUSE(X,Y)" as *implicit instances*. For "CAUSE(a,b)" to be inferred by this rule, "CAN_CAUSE(a,b)" must be written in single sentences so that TextRunner can recognize it. This means that such instances are SS instances.

Actually their algorithm was tuned to prefer SS instances and around 70% of the acquired valid instances were in this category. Certainly, their method uses more complex rules that can infer NS instances, but the precision of the instances inferred by such non-paraphrase rules is quite low (around 20%) [1]. Also they have not empirically examined how many NS instances are actually found in their output.

In contrast, we focus on more complex rules that can infer NS instances. An important point is that their method cannot deal with considerable parts of our complex rules because their inference algorithm for markov logic network (Huynh and Mooney, 2008) poses some restrictions on recursive rules as discussed in Section 4.1. We also empirically show that their scoring mechanism does not lead to higher precision at least in our setting, although we have not checked whether this is due to the restrictions on recursive rules or is due to some other reasons.

Another work in this category, Carlson et al. (2010) hypothesized instances using Horn clauses discovered by Inductive Logic Programming (Quinlan and Cameron-Jones, 1993). Tsuchida et al. (2010) generates hypothesized relation instances by substituting words in seed instances with distributionally similar words.

In bioinformatics, there are attempts to develop methods to help discoveries of relation instances by linking clues from different literatures (Swanson, 1986; Srinivasan, 2004; Hu et al., 2006; Hristovski et al., 2008). These methods cannot be easily adapted to other domains, because they require heavily engineered resources like databases of MEDLINE records and hand-annotation with MeSH [2] metadata. These also require some input and/or interactions to capture the interests of the human expert. Our aim is to infer unseen NS instances without such resources and human effort.

Recently, De Saeger et al. (2011) have shown that it is possible to acquire SS instances from highly complex and infrequent expressions, using word classes and lexico-syntactic pattern fragments, which they call *partial patterns*. Such approach may prove useful for acquiring NS instances too, as the method can acquire relation instances without considering any pattern connecting the two words of the instance. Yet their work focused only on SS instances.

3 Proposed Method

Our method takes a set of seed relation instances and a large corpus as input. The seed instances are obtained using De Saeger et al. (2009), after some rudimentary cleaning (Section 4). Then, our method induces possible inference rules, and assigns scores to the instances inferred by the rules. The high-ranked instances are provided as output.

[1] They acquired a total 2.6M instances with 50% precision for a variety of relation types. Also, about 1.25M SS instances in those were inferred by simple paraphrase rules with 80% precision. The precision of the instances inferred by non-paraphrase rules can be estimated as the solution for $0.50 = \frac{1.25 \times 0.80 + (2.6 - 1.25) \times x}{2.6}$ as x $(= 0.22)$.

[2] http://www.nlm.nih.gov/mesh/

In the following, we describe the details of each step. Note that the algorithm presented here can produce SS instances as well as NS instances. In the evaluation, *potential* SS instances are excluded from the output by checking the co-occurrence of the word pairs in the instances in single sentences.

3.1 Inducing Inference Rules

We consider inference rules (a special case of Horn-clauses) that have the following form:

$$\text{“X } \textit{pattern} \text{ Y”} \wedge \mathrm{R}_{\mathrm{SEED}}(\mathrm{Y,Z}) \rightarrow \mathrm{R}_{\mathrm{HYPO}}(\mathrm{X,Z})$$

This rule hypothesizes relation instances by substituting the first argument Y of a seed instance with X where the connection between X and Y is provided by “X *pattern* Y”. [3] To handle multi-word expressions such as “*red wine*”, we allow NP chunks to fill the slots of the X, Y or Z variables. Our motivation here is to have a more exhaustive set of relation instances for a given relation type. As a starting point we focus on the simple rules as the one above, although a wide range of other forms are possible. Sharing a variable (i.e. Y) guarantees that at least some relationship between X and Z holds and reduces the risk of generating meaningless rules, while the arbitrariness of the pattern allows us to consider a broad range of evidence to find new instances.

For inducing inference rules, first we find two seed instances that have the same second argument. Suppose that the target relation is causality, and we have two seed instances, $\mathrm{CAUSE}(\textit{bronchitis}, \textit{cough})$ and $\mathrm{CAUSE}(\textit{mold}, \textit{cough})$, with common second argument “*cough*”. Then, we can search for “X *pattern* Y”, which informs us of a certain relation between the first arguments of the two seed instances, *bronchitis* and *mold* in this example. We may be able to find “X causes Y” from the expression “.. *mold* causes *bronchitis* ..” and induce the following rule that infers one seed instance from another seed instance.

$$\text{”}\textit{mold}\text{ causes }\textit{bronchitis}\text{”} \wedge \mathrm{CAUSE}_{\mathrm{SEED}}(\textit{bronchitis},\textit{cough}) \rightarrow \mathrm{CAUSE}_{\mathrm{HYPO}}(\textit{mold}, \textit{cough})$$

By generalizing *mold*, *bronchitis*, and *cough* to variables X, Y and Z, we can obtain the following acceptable rule that can be interpreted as “if X is a cause of a cause (Y) of Z, X is a cause of Z”.

$$\text{“X causes Y”} \wedge \mathrm{CAUSE}_{\mathrm{SEED}}(\mathrm{Y,Z}) \rightarrow \mathrm{CAUSE}_{\mathrm{HYPO}}(\mathrm{X,Z})$$

[3] We also consider the rules where Z appears as the first arguments, i.e., “X *pattern* Y” $\wedge$ $\mathrm{R}_{\mathrm{SEED}}$(Z,Y) $\rightarrow$ $\mathrm{R}_{\mathrm{HYPO}}$(Z,X).

Here, by assuming that $\mathrm{CAUSE}_{\mathrm{HYPO}}$ (*bronchitis*, *cough*) is inferred from $\mathrm{CAUSE}_{\mathrm{SEED}}$ (*mold*, *cough*), we also obtain the following unacceptable rule, which contradicts with common sense.

$$\text{“Y causes X”} \wedge \mathrm{CAUSE}_{\mathrm{SEED}}(\mathrm{Y,Z}) \rightarrow \mathrm{CAUSE}_{\mathrm{HYPO}}(\mathrm{X,Z})$$

We can say that the former is a *swapped rule* of the latter and vice-versa. The above rule can be interpreted as “If Y is a cause of X and Z, then X is a cause of Z”, which is nonsense. In Section 3.2.1, we introduce a heuristic to remove such unacceptable swapped rules.

To alleviate pattern ambiguity, all the patterns in the rules are *class dependent patterns* (Section 2). The induced rules are augmented with class restrictions on the variables in the pattern:

$$\text{“X:}\texttt{virus}\text{ causes Y:}\texttt{sickness}\text{”} \wedge \mathrm{CAUSE}_{\mathrm{SEED}}(\mathrm{Y,Z}) \rightarrow \mathrm{CAUSE}_{\mathrm{HYPO}}(\mathrm{X,Z})$$

As word classes for the class restrictions, we used a word clustering result obtained by applying a probabilistic word clustering algorithm (Kazama and Torisawa, 2008) to dependency relations extracted from 100M Web pages. The results are represented by the probability distribution $P(c|w)$ where w is a word (the number of words is 1M in this paper) and c is a class identifier, which is a hidden variable in a predefined set C.

To have a *discrete* boundary of class membership, we assume w belongs to class c if and only if $P(c|n) \geq 0.2$ or $c = \arg\max_{c \in C} P(c|n)$. We set $|C|$ to 500, following De Saeger et al. (2009). Also, the classes are just class identifiers (i.e., integers). To simplify the explanation, we omit these class restrictions in the rest of our paper.

Finally, to discard unproductive rules, all the rules that cannot generate more than M seed relation instances using the other seed instances are discarded. In this work, we set M to 10.

3.2 Scoring Hypothesized Instances

After inducing the rules, we hypothesize relation instances by applying the rules to an input corpus, and assign scores to the instances. *Instance score* for a hypothesized instance is defined as the sum of the scores of the rules that generated the instance. Rule score $r\_score(r)$ is an approximated precision of the instances inferred by the rule, assuming that the seed instances are *correct* relations and the others are not.

$$r\_score(r) = \frac{\text{\# of seeds in hypotheses by } r}{\text{\# of hypotheses by } r}$$

Here, r is an inference rule.

Instance score h_score is defined as the sum of the applicable rule scores.

$$h\_score(h) = \sum_{r \in \mathbf{Irules}(h, \mathbf{Seeds}, \mathbf{Rules})} r\_score(r)$$

Here, h is a hypothesized instance, and **Rules** is a set of inference rules. **Seeds** is a set of seed relation instances, and **Irules** is a set of rules in **Rules** that infer h from **Seeds**.

We use the following additional heuristics during this ranking scheme.

3.2.1 Removing unacceptable swapped rules

The first heuristic is introduced for removing unacceptable swapped rules. Recall that our inference rule always has its swapped rule like the following two example rules.

A "X causes Y" $\wedge$ $\text{CAUSE}_{\text{SEED}}$ (Y, Z) $\rightarrow$ $\text{CAUSE}_{\text{HYPO}}$ (X, Z)

B "Y causes X" $\wedge$ $\text{CAUSE}_{\text{SEED}}$ (Y, Z) $\rightarrow$ $\text{CAUSE}_{\text{HYPO}}$ (X, Z)

Here, we have three observations: 1) if one rule is acceptable, its swapped rule is not (e.g., rule A is acceptable but rule B is not), 2) if a rule often generates reflexive relation instances, i.e., R_{HYPO} (x,x) for some word "x", the rule is likely to be unacceptable, and 3) the swapped rule of such an unacceptable rule (like rule B) often represents a transitive relation. For example, above rule A represents a transitive relation for causality (X $\rightarrow$ Y $\rightarrow$ Z then X $\rightarrow$ Z). On the other hand, the seed instance and the pattern in rule B represent that the same cause Y results in X and Z. If the seed instance and the pattern represents the same causal relation instance, $\text{CAUSE}_{\text{HYPO}}$ (X,Z) becomes $\text{CAUSE}_{\text{HYPO}}$ (x,x).

We found that rules generating many R_{HYPO} (x,x) attain a high score in our rule scoring. To remedy this we set $r\_score(r)$ to 0 if r generates more R_{HYPO} (x,x) than its swapped version. If the two rules do not generate R_{HYPO} (x,x) or the number of R_{HYPO} (x,x) by the two rules are the same, this is not applied. We refer to this as (X,X)-based rule filtering hereafter.

3.2.2 Excluding highly vague words

Our inference rules are based on *pivot* words, i.e., the shared variable Y in "X pattern Y $\wedge$ R_{SEED}(Y,Z) $\rightarrow$ R_{HYPO}(X,Z)". If the pivot is so vague that it likely refers to different objects in the seed and the pattern, the inference would be wrong (Schoenmackers et al., 2010). For example, we can generate false hypothesis $\text{CAUSE}_{\text{HYPO}}$ (*cerebral stroke*, *pneumonia*) from "*cerebral stroke* causes *disease*" and $\text{CAUSE}_{\text{SEED}}$ (*disease*, *pneumonia*). Here "*disease*" is highly vague, and likely refers to different types of diseases in different texts.

To address this problem we prepared a stop word list and discarded relation instances that contain one of these stop words. We calculate the document frequency of each word in the 600M page Web corpus, and regard a word whose document frequency is higher than threshold T as a stop word. We set T to 400,000 and obtain about 15,000 stop words in our experiments.

4 Evaluation

Our evaluation focuses on three main questions: 1) How accurate are the *NS instances* hypothesized by our method? 2) How accurate are all the instances hypothesized by our method? 3) Does our method infer relation instances with higher precision than competing methods? After explaining the experimental setting we answer each question from our experimental results. We also analyze the cause of errors at the end of this section.

4.1 Experimental Setting

We evaluated our results on the three relations:

***Causality*(X,Y):** X can directly or indirectly cause Y. e.g., CAUSE(*allergen*, *allergy*).

***Prevention*(X,Y):** X can directly or indirectly help to avoid the occurrence of Y. e.g., PREVENTION(*coffee*, *drowsiness*).

***Material*(X,Y)** X is a material or ingredient of Y. e.g., MATERIAL(*grape*, *wine*).

We employed the evaluation scheme introduced in Tsuchida et al. (2010). To evaluate the correctness of a given hypothesized instance, we presented three human judges with short texts as possible evidence. For each hypothesized instance, we collected up to 20 short texts from the search results provided by Yahoo! API [4] for a query consisting of the two nouns and a priming term for each target relation, i.e., the Japanese word for "*cause*", "*material*" or "*prevention*". All texts were collected from February to the middle of March in 2011.

The expectation behind this scheme is that many of correct *NS instances* in a smaller corpus,

[4] http://developer.yahoo.co.jp/webapi/search/

i.e., the 600M page Web corpus, can be found with explicit evidence in a larger corpus, i.e., pages accessible through a commercial search engine.

In our evaluation scheme, our three annotators mark a hypothesized instance as correct if they find "sufficient evidence" (see below) in at least one of the presented text snippets. We say a text snippet contains "sufficient evidence" if it either explicitly asserts the target relation between the word pair, or implicitly presupposes it. A hypothesized instance is judged incorrect when 1) the provided texts do not present sufficient evidence, or 2) a relation instance is not informative enough without further context (e.g., CAUSE(*insulin, change*), we don't know what *change* can be caused by *insulin* without further context). Correctness of a hypothesized instance is determined by the judges' majority vote. The inter-rater agreement (Fleiss' kappa) was 0.57 for causality, 0.56 for prevention and 0.57 for material, indicating the judgements are reasonably stable.

For each relation, the seed instances were the top 20,000 instances given by our implementation of De Saeger et al. (2009) with the class-based cleaning that took about 15 minutes. More precisely, we discarded inappropriate instances by manually identifying and removing semantic classes that are inappropriate for the target relation. The precision of the seed instances was 81% for *causality* , 78% for *material* and 44% for *prevention*. We used about 25 seed patterns for each relation, which were created by one of the authors with one hour tuning. We show some examples of the seed patterns (translated from Japanese).

***Causality*(X,Y)** "X causes Y", "Y caused by X", "X that causes Y", ...

***Prevention*(X,Y)** "X prevents Y", "Y prevented by X", "X that prevents Y", ...

***Material*(X,Y)** "X is material of Y", "Y made from X", "X that is material of Y", ...

From the seed instances, our method induced 24,044 rules for causality, 17,868 for prevention and 14,978 for material, and obtained 3.04M hypothesized instances for causality, 2.44M for prevention and 2.17M for material. These hypothesized instances do not include the seed instances. Table 1 shows some hypothesized instances.

In our experiments we compared three systems.

SUM: Proposed method.

MLN: Markov logic network based scoring method of Schoenmackers et al. (2010).

MLN(X,X): MLN with the weight of the rules removed by (X,X)-based rule filtering set to 0.

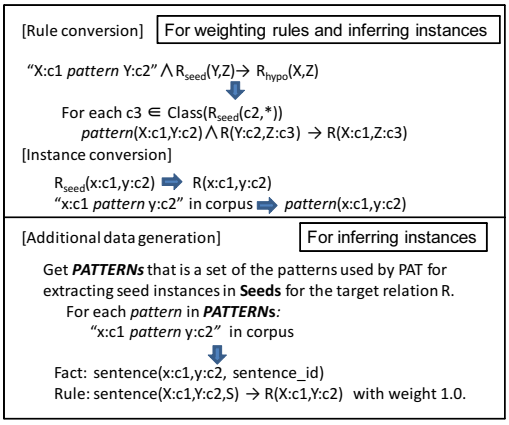

Figure 1: Data generation for MLN. X, Y and Z are variables, x, y and z are nouns. c1, c2 and c3 are noun classes, and Class(R_{SEED} $(c, *)$) is a set of classes, c_i, for which some R_{SEED} (y:c, z:c_i) exists. "x:c" indicates that x belongs to class c, and "X:c" means that X restricts possible nouns by class c.

We do not compare our method to Carlson et al. (2010), since Schoenmackers et al. (2010) already showed that MLN outperforms that method.

In preparing MLN, we converted our rules to the format used in Schoenmackers et al. (2010) as shown in Figure 1. Unlike our method, all rule variables in their work have class (or hypernym) restrictions, and predicates with different argument classes are treated as different. Truly recursive rules that have the same predicate with the same class restrictions in both head and body were removed due to limitations in their inference algorithm.[5] We removed 10,944 rules for causality (46% of all rules), 8,614 (48%) for prevention and 7,074 (47%) for material, respectively. When inferring instances, we also generated the additional rules and facts (lower half of Figure 1) that reflect their assumption that an instance frequently extracted from single sentences is likely to be true. Also, we empirically set the standard deviations of the Gaussian prior for learning rule weights and the prior weight of all unknown facts for inference to 2.0 and -3.0. We did not use the variable Gaussian prior to discount the weights of longer rules, because the lengths of all our rules are the same.

4.2 Evaluation Results

For SUM, MLN and MLN(X,X), we investigated the precision of the top 10,000 *NS instances* —

[5] The freely available inference system for markov logic Alchemy (`http://alchemy.cs.washington.edu/`) allows recursive rules, but turned out to be too slow for practical use. In our experiments we gave up on it after waiting more than a week for the program to finish.

Table 1: Examples of evaluated hypothesized instances and the rules that inferred the instances (translated from Japanese). The meaning of "N4S" is described in Section 4.2. All the instances marked with"NS", "N4S" and "SS" were judged as *correct*. "*UR" marks incorrect instances generated by *unacceptable rules*. "*AR" denotes incorrect instances generated by *acceptable rules*. "*NE" denotes instances having *no evidence* in the provided texts, though we expected that the instances may be correct based on their original rules and seed instances. For simplicity, we omit the class restrictions of rules.

	Label	Inferred instance	Samples of rules that generated the left hypothesized instance.
Causality	N4S	CAUSE$_{\text{HYPO}}$ (Z=*SO2 gas*, X=*allergy symptoms*), Y= *asthma*	CAUSE$_{\text{HYPO}}$ (Z,X) ← **X** *is caused by* **Y** ∧ CAUSE$_{\text{SEED}}$ (Z,Y) CAUSE$_{\text{HYPO}}$ (Z,X) ← **X** *gets worse by* **Y** ∧ CAUSE$_{\text{SEED}}$ (Z,Y)
	NS	CAUSE$_{\text{HYPO}}$ (X=*reactive oxygen species*, Z=*aneurysm*), Y=*high blood pressure*	CAUSE$_{\text{HYPO}}$ (X,Z) ← **X**'*s increase causes* **Y** ∧ CAUSE$_{\text{SEED}}$ (Y,Z) CAUSE$_{\text{HYPO}}$ (X,Z) ← **X** *is a cause of* **Y** ∧ CAUSE$_{\text{SEED}}$ (Y,Z)
	SS	CAUSE$_{\text{HYPO}}$ (X=*pesticide*, Z=*colorectal cancer*), Y=*harmful substance*	CAUSE$_{\text{HYPO}}$ (X,Z) ← **Y** *is contained in* **X** ∧ CAUSE$_{\text{SEED}}$ (Y,Z) CAUSE$_{\text{HYPO}}$ (X,Z) ← **X** *is called* **Y** ∧ CAUSE$_{\text{SEED}}$ (Y,Z)
	*UR	CAUSE$_{\text{HYPO}}$ (X=*bilirubin*, Z=*colorectal cancer*) Y=*bile*	CAUSE$_{\text{HYPO}}$ (X,Z) ← **X** *is contained in* **Y** ∧ CAUSE$_{\text{SEED}}$ (Y,Z) CAUSE$_{\text{HYPO}}$ (X,Z) ← **X** *is excreted in* **Y** ∧ CAUSE$_{\text{SEED}}$ (Y,Z)
	*AR	CAUSE$_{\text{HYPO}}$ (Z=*potato crisp*, X=*atherosclerosis*) Y=*lifestyle diseases*	CAUSE$_{\text{HYPO}}$ (Z,X) ← **Y** *is a cause of* **X** ∧ CAUSE$_{\text{SEED}}$ (Z,Y) CAUSE$_{\text{HYPO}}$ (Z,X) ← **X** *caused by* **Y** ∧ CAUSE$_{\text{SEED}}$ (Z,Y)
	*AR	CAUSE$_{\text{HYPO}}$ (X=*tobacco*, Z=*food poisoning*) Y=*harmful component*	CAUSE$_{\text{HYPO}}$ (X,Z) ← **Y** *contained in* **X** ∧ CAUSE$_{\text{SEED}}$ (Y,Z) CAUSE$_{\text{HYPO}}$ (X,Z) ← **X** *contains* **Y** ∧ CAUSE$_{\text{SEED}}$ (Y,Z)
	*NE	CAUSE$_{\text{HYPO}}$ (Z=*magnesium chloride*, X=*atherosclerosis*) Y=*high blood pressure*	CAUSE$_{\text{HYPO}}$ (Z,X) ← **X** *gets worse by* **Y** ∧ CAUSE$_{\text{SEED}}$ (Z,Y) CAUSE$_{\text{HYPO}}$ (Z,X) ← **Y** *triggers* **X** ∧ CAUSE$_{\text{SEED}}$ (Z,Y)
Prevention	N4S	PREVENTION$_{\text{HYPO}}$ (X=*blue-backed fish*, Z=*cerebral thrombosis*), Y=*eicosapentaenoic acid*	PREVENTION$_{\text{HYPO}}$ (X,Z) ← **Y** *is extracted from* **X** ∧ PREVENTION$_{\text{SEED}}$ (Y,Z) PREVENTION$_{\text{HYPO}}$ (X,Z) ← **X** *is rich in* **Y** ∧ PREVENTION$_{\text{SEED}}$ (Y,Z)
	NS	PREVENTION$_{\text{HYPO}}$ (Z=*sunflower oil*,X=*heart disease*), Y=*high blood pressure*, Y1=*linoleic acid*	PREVENTION$_{\text{HYPO}}$ (Z,X) ← **X** *is caused by* **Y** ∧ PREVENTION$_{\text{SEED}}$ (Z,Y) PREVENTION$_{\text{HYPO}}$ (Z,X) ← **Y1** *is contained in* **Z** ∧ PREVENTION$_{\text{SEED}}$ (Y1,X)
	SS	PREVENTION$_{\text{HYPO}}$ (X=*sesame lignan*, Z=*cerebral stroke*), Y=*antioxidant effects*,Y1=*high blood pressure*	PREVENTION$_{\text{HYPO}}$ (X,Z) ← **X** *having* **Y** ∧ PREVENTION$_{\text{SEED}}$ (Y,Z) PREVENTION$_{\text{HYPO}}$ (X,Z) ← **Y1** *gives rise to* **Z** ∧ PREVENTION$_{\text{SEED}}$ (X,Y1)
	*UR	PREVENTION$_{\text{HYPO}}$ (Z=*niacin*, X=*kidney disease*), Y=*high blood pressure*	PREVENTION$_{\text{HYPO}}$ (Z,X) ← **Y** *is accompanied by* **X** ∧ PREVENTION$_{\text{SEED}}$ (Z,Y) PREVENTION$_{\text{HYPO}}$ (Z,X) ← **X** *is a cause of* **Y** ∧ PREVENTION$_{\text{SEED}}$ (Z,Y)
	*NE	PREVENTION$_{\text{HYPO}}$ (Z=*egg*, X=*dizziness*), Y=*anemia*, Y1=*iron*	PREVENTION$_{\text{HYPO}}$ (Z,X) ← **X** *is caused by* **Y** ∧ PREVENTION$_{\text{SEED}}$ (Z,Y) PREVENTION$_{\text{HYPO}}$ (Z,X) ← **Z** *contains* **Y1** ∧ PREVENTION$_{\text{SEED}}$ (Y1,X)
Material	N4S	MATERIAL$_{\text{HYPO}}$ (X=*red grape*, Z=*cuvee*), Y=*pinot noir*	MATERIAL$_{\text{HYPO}}$ (X,Z) ← **X** *such as* **Y** ∧ MATERIAL$_{\text{SEED}}$ (Y,Z) MATERIAL$_{\text{HYPO}}$ (X,Z) ← **X** *including* **Y** ∧ MATERIAL$_{\text{SEED}}$ (Y,Z)
	NS	MATERIAL$_{\text{HYPO}}$ (Z=*sugar beet*, X=*hydrogen*), Y=*ethanol*	MATERIAL$_{\text{HYPO}}$ (Z,X) ← **X** *is extracted from* **Y** ∧ MATERIAL$_{\text{SEED}}$ (Z,Y) MATERIAL$_{\text{HYPO}}$ (Z,X) ← **Y** *is converted into* **X** ∧ MATERIAL$_{\text{SEED}}$ (Z,Y)
	SS	MATERIAL$_{\text{HYPO}}$ (Z=*corn*, X=*ethylene*), Y=*ethanol*	MATERIAL$_{\text{HYPO}}$ (Z,X) ← **X** *is made from* **Y** ∧ MATERIAL$_{\text{SEED}}$ (Z,Y) MATERIAL$_{\text{HYPO}}$ (Z,X) ← **Y** *is material of* **X** ∧ MATERIAL$_{\text{SEED}}$ (Z,Y)
	*UR	MATERIAL$_{\text{HYPO}}$ (Z=*blueberry*, X= *plain yogurt*), Y=*blueberry jelly*	MATERIAL$_{\text{HYPO}}$ (Z,X) ← **Y** *is mixed in* **X** ∧ MATERIAL$_{\text{SEED}}$ (Z,Y) MATERIAL$_{\text{HYPO}}$ (Z,X) ← *put* **Y** *in* **X** ∧ MATERIAL$_{\text{SEED}}$ (Z,Y)
	*AR	MATERIAL$_{\text{HYPO}}$ (X=*sugarcane*, Z= *zero-emisions vehicle*), Y=*ethanol*	MATERIAL$_{\text{HYPO}}$ (X,Z) ← **Y** *extracted from* **X** ∧ MATERIAL$_{\text{SEED}}$ (Y,Z) MATERIAL$_{\text{HYPO}}$ (X,Z) ← **Y** *made from* **X** ∧ MATERIAL$_{\text{SEED}}$ (Y,Z)

Table 2: Results from the binomial one-tailed test between SUM and the other methods. The significance level is 0.05. For each relation, cells show the number of data points (8 max) in the precision graph (like Figure 3) where "SUM wins / SUM loses / no significant difference".

	MLN(X,X)	MLN	MAX	No (X,X) filtering	No rule score
Causality	0 / 0 / 8	6 / 0 / 2	8 / 0 / 0	6 / 0 / 2	3 / 0 / 5
Prevention	8 / 0 / 0	7 / 0 / 1	6 / 0 / 2	0 / 0 / 8	0 / 0 / 8
Material	7 / 0 / 1	7 / 0 / 1	1 / 0 / 7	8 / 0 / 0	2 / 0 / 6

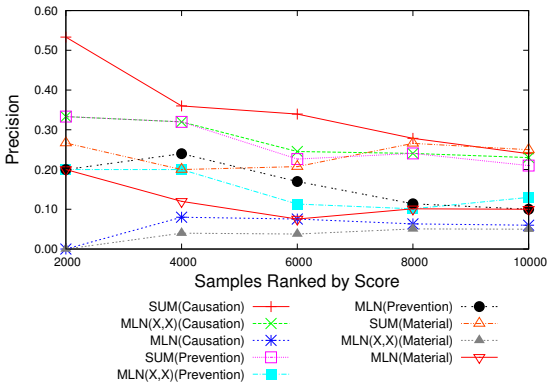

Figure 2: Precision of the top 10,000 NS instances.

word pairs that do not co-occur in any single sentence — using 100 random samples. The precision curves in Figure 2 show that SUM outperforms both MLN and MLN(X,X). Although the precision of the top 10,000 NS instances obtained by SUM is relatively low (20% to 30%), we do think this is a promising result given the difficulty of the task. In tasks such as innovation support and risk management, one must explore the border between the known and unknown. We expect even hypothesized instances with 20 to 30% precision can be useful in such contexts.

Next, we would like to address the question of how many acquired NS instances are not written in our corpus at all. Answering this question precisely is difficult because of anaphora and ellipsis. Therefore we assume that if the two nouns of an NS instance do not co-occur in any four sentence window ("N4S") in the corpus, the instance is unlikely to be mentioned explicitly in any form in the corpus. For causality, 44 of the 100 evaluated samples were N4S instances, 8 of which were correct

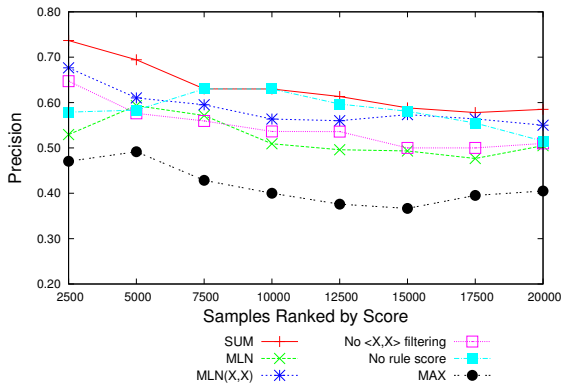

Figure 3: Causality: top 20,000 results' precision

(18% precision). The evaluated prevention and material samples respectively contained 22 and 35 N4S instances, with 2 and 8 correct ones (9% and 23% precision). We expect that at least some of these instances are not mentioned in our corpus.

To confirm the superiority of the proposed method, we investigated the precision of the top 20,000 instances of each method without removing SS instances from the evaluation data using 200 random samples. Our method showed 59% for causality (Fig. 3), 46% for prevention and 43% for material of the top 20,000 in precision, which is considerably higher than the precision of the NS samples only.

In addition to MLN and MLN(X,X) we also evaluated the following three baseline methods to confirm that all the design choices effectively contribute to the performance of our method.

No (X,X) filtering: SUM without (X,X)-based filtering.
No rule score: SUM, where all rule scores are constant.
MAX: A variant of the instance scoring function,

$$\max_{r \in \mathbf{Irules}(h,\mathbf{Seeds},\mathbf{Rules})} r\_score(r).$$

Figure 3 shows the precision curves for causality. In all relations SUM outperforms MLN, MLN(X,X) and all baseline methods. Table 2 shows the results of the binomial one-tailed test between SUM and each compared method, and suggests that SUM had statistically significant improvements in many cases. SUM showed only minor improvements compared with "No rule score". This suggests that the instances inferred by *many* rules tend to be correct, irrespective of the rule scores. We think this is due to the overall quality of the rules, as this would not be the case if many rules were invalid. Table 2 also shows that for transitive relations (causality and material), (X,X)-based rule filtering is effective.

4.3 Error Analysis

Table 1 shows some instances hypothesized by the proposed method. We distinguish between incorrect instances generated by unacceptable rules (marked "*UR" in Table 1) and others. We found that about 60% of incorrect instances are generated by unacceptable rules. For example, "$\text{CAUSE}_{\text{HYPO}}$ (X=*bilirubin*, Z=*colorectal cancer*)" was generated by the rule "$\text{CAUSE}_{\text{HYPO}}$ (X,Z) $\leftarrow$ X *is contained in* Y $\wedge$ $\text{CAUSE}_{\text{SEED}}$ (Y,Z)".

Incorrect instances generated by seemingly correct rules are marked with "*AR" in Table 1. These can further be categorized into three types (examples of each are in Table 1).

Type A: Uninformative instances for our evaluation criteria (See Section 4.1).
e.g., $\text{CAUSE}_{\text{HYPO}}$ (Z=*potato crisps*, X=*atherosclerosis*). The validity of this instance depends on context, i.e., the amount of potato crisps that one eats.

Type B: Instances generated with vague pivot words (See Section 3.2.2).
e.g., $\text{CAUSE}_{\text{HYPO}}$ (X=*tobacco*, Z=*food poisoning*), Y=*harmful component*. *harmful component* is vague.

Type C: Instances generated from incorrect seed instances.
e.g., $\text{MATERIAL}_{\text{HYPO}}$ (X=*sugarcane*, Z=*zero-emissions vehicle*), derived from the incorrect seed instance $\text{MATERIAL}_{\text{SEED}}$ (Y=*ethanol*, Z=*zero-emisions vehicle*).

The ratio was roughly 10% for type **A**, 5% for type **B** and 5% for type **C** for all the relation types. For the remaining incorrect instances (about 20%), the judges could not find sufficient evidence in the presented text snippets, although some such instances do appear to be valid. Such instances are marked "*NE". This suggests our evaluation scheme may be underestimating the true precision.

5 Conclusion

This paper introduced an inference-based method that takes a set of seed relation instances as input, and outputs hypothesized instances using inference rules induced from these seed instances.

We showed that our method can infer valid relation instances whose component nouns do not co-occur in any single sentence or any four sentence window even in a 600M page Web corpus. We expect this result is promising for inferring new knowledge, because such instances may contain instances not mentioned in the context of that particular semantic relation even in 600M Web pages.

Acknowledgments

The authors thank Dr. Tuyen N. Huynh of SRI International, who kindly provided the source code of Huynh and Mooney (2008) for our experiments.

References

Eugene Agichtein and Gravano Luis. 2001. Snowball: Extracting Relations from Large Plain-Text Collections. In *Proc. of the 5th ICDL*, pages 85–94.

Michele Banko and Oren Etzioni. 2008. The tradeoffs between open and traditional relation extraction. In *Proc. of the 46th ACL-08:HLT*, pages 28–36.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proc of the 24th AAAI*, pages 1306–1313.

Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata. 2009. Large Scale Relation Acquisition Using Class Dependent Patterns. In *Proc. of the 9th ICDM*, pages 764–769.

Stijn De Saeger, Kentaro Torisawa, Masaaki Tsuchida, Junfichi Kazama, Chikara Hashimoto, Ichiro Yamada, Jong-Hoon Oh, Istvan Varga, and Yulan Yan. 2011. Relation Acquisition using Word Classes and Partial Patterns. In *Proc. of the EMNLP2011*, pages 825–835.

Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2004. Web-Scale Information Extraction in Knowitall (Preliminary Results). In *Proc. of the 13th WWW*, pages 100–110.

D. Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin. 2008. Literature-Based Knowledge Discovery using Natural Language Processing. *Literature-based Discovery, Information Science and Knowledge Management*, 15:133–152.

Xiaouha Hu, Xiaodan Zhang, Illhoi Yoo, and Yanqing Zhang. 2006. A semantic approach for mining hidden links from complementary and non-interactive biomedical literature. In *Proc. of the 6th SDM*, pages 200–209.

Tuyen N. Huynh and Raymond J. Mooney. 2008. Discriminative structure and parameter learning for markov logic networks. In *Proc. of the 25th ICML*, pages 416–423.

Jun'ichi Kazama and Kentaro Torisawa. 2008. Inducing Gazetteers for Named Entity Recognition by Large-scale Clustering of Dependency Relations. In *Proc. of the 46th ACL*, pages 407–415.

Marius Paşca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain. 2006. Names and Similarities on the Web: Fact Extraction in the Fast Lane. In *Proc. of the COLING-ACL06*, pages 809–816.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proc. of the COLING-ACL06*, pages 113–120.

J. R. Quinlan and R. M. Cameron-Jones. 1993. FOIL: A Midterm Report. In *Proc. of the ECML*, pages 3–20.

Matthew Richardson and Pedro Domingo. 2006. Markov logic networks. *Machine Learning*, 26:107–136.

Stefan Schoenmackers, Oren Etzioni, Daniel S. Weld, and Jesse Davis. 2010. Learning first-order horn clauses from web text. In *Proc. of EMNLP2010*, pages 1088–1098.

Padmini Srinivasan. 2004. Text mining: generating hypotheses from medline. *Journal of the American Society for Information Science and Technology*, 55(5):396–413.

Don R. Swanson. 1986. Undiscovered public knowledge. *Library Quarterly*, 56(2):103–118.

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

Masaaki Tsuchida, Stijn De Saeger, Kentaro Torisawa, Masaki Murata, Jun'ichi Kazama, Kow Kuroda, and Hayato Ohwada. 2010. Large scale similarity-based relation exapnsion. In *Proc of the 4th IUCS*, pages 140–147.