Acquiring Inference Rules with Temporal Constraints by Using Japanese Coordinated Sentences and Noun-Verb Co-occurrences

Kentaro Torisawa
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi-shi, Ishikawa-ken, 923-1211 JAPAN
torisawa@jaist.ac.jp

Abstract

This paper shows that inference rules with temporal constraints can be acquired by using verb-verb co-occurrences in Japanese coordinated sentences and verb-noun co-occurrences. For example, our unsupervised acquisition method could obtain the inference rule "If someone enforces a law, usually someone enacts the law at the same time as or before the enforcing of the law" since the verbs "enact" and "enforce" frequently co-occurred in coordinated sentences and the verbs also frequently co-occurred with the noun "law". We also show that the accuracy of the acquisition is improved by using the occurrence frequency of a single verb, which we assume indicates how *generic* the meaning of the verb is.

1 Introduction

Our goal is to develop an unsupervised method for acquiring inference rules that describe logical implications between event occurrences. As clues to find the rules, we chose Japanese coordinated sentences, which typically report two events that occur in a certain temporal order. Of course, not every coordinated sentence necessarily expresses implications. We found, though, that reliable rules can be acquired by looking at co-occurrence frequencies between verbs in coordinated sentences and co-occurrences between verbs and nouns. For example, our method could obtain the rule "If someone enforces a law, usually someone enacts the law at the same time as or before the enforcing of the law". In our experiments, when our method produced 400 rules for 1,000 given nouns, 70% of the rules were considered proper by at least three of four human judges.

Note that the acquired inference rules pose temporal constraints on occurrences of the events described in the rules. In the "enacting-and-enforcing-law" example, the constraints were expressed by the phrase "at the same time as or before the event of". We think such temporally constrained rules should be beneficial in various types of NLP applications. The rules should allow Q&A systems to *guess* or *restrict* the time at which a certain event occurs even if they cannot directly find the time in given documents. In addition, we found that a large part of the acquired rules can be regarded as *paraphrases*, and many possible applications of paraphrases should also be target applications.

To acquire rules, our method uses a score, which is basically an approximation of the probability that particular coordinated sentences will be observed. However, it is weighted by a *bias*, which embodies our assumption that frequently observed verbs are likely to appear as the consequence of a proper inference rule. This is based on our intuition that frequently appearing verbs have a *generic* meaning and tend to describe a wide range of situations, and that natural language expressions referring to a wide range of situations are more likely to be a consequence of a proper rule than *specific* expressions describing only a narrow range of events. A similar idea relying on word co-occurrence was proposed by Geffet and Dagan (Geffet and Dagan, 2005) but our method is simpler and we expect it to be applicable to a wider range of vocabularies.

Research on the automatic acquisition of inference rules, paraphrases and entailments has received much attention. Previous attempts have used, for instance, the similarities between case frames (Lin and Pan-

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 57–64,
New York, June 2006. ©2006 Association for Computational Linguistics

tel, 2001), anchor words (Barzilay and Lee, 2003; Shinyama et al., 2002; Szepektor et al., 2004), and a web-based method (Szepektor et al., 2004; Geffet and Dagan, 2005). There is also a workshop devoted to this task (Dagan et al., 2005). The obtained accuracies have still been low, however, and we think searching for other clues, such as coordinated sentences and the bias we have just mentioned, is necessary. In addition, research has also been done on the acquisition of the temporal relations (Fujiki et al., 2003; Chklovski and Pantel, 2004) by using coordinated sentences as we did, but these works did not consider the *implications* between events.

2 Algorithm with a Simplified Score

In the following, we begin by providing an overview of our algorithm. We specify the basic steps in the algorithm and the form of the rules to be acquired. We also examine the direction of implications and temporal ordering described by the rules. After that, we describe a simplified version of the scoring function that our algorithm uses and then discuss a problem related to it. The bias mechanism, which we mentioned in the introduction, is described in the section after that.

2.1 Procedure and Generated Inference Rules

Our algorithm is given a noun as its input and produces a set of inference rules. A produced rule expresses an implication relation between two descriptions including the noun. Our basic assumptions for the acquisition can be stated as follows.

- If verbs v_1 and v_2 frequently co-occur in coordinated sentences, the verbs refer to two events that actually frequently co-occur in the real world, and a sentence including v_1 and another sentence including v_2 are good candidates to be descriptions that have an implication relation and a particular temporal order between them.

- The above tendency becomes stronger when the verbs frequently co-occur with a given noun n; i.e., if v_1 and v_2 frequently co-occur in coordinated sentences and the verbs also frequently co-occur with a noun n, a sentence including v_1 and n and another sentence including v_2 and n are good candidates to be descriptions that have an implication relation between them.

Our procedure consists of the following steps.

Step 1 Select M verbs that take a given noun n as their argument most frequently.

Step 2 For each possible pair of the selected verbs, compute the value of a scoring function that embodies our assumptions, and select the N verb pairs that have the largest score values. Note that we exclude the combination of the same verb from the pairs to be considered.

Step 3 If the score value for a verb pair is higher than a threshold θ and the verbs take n as their syntactic objects, generate an inference rule from the verb pair and the noun.

Note that we used 500 as the value of M. N was set to 4 and θ was set to various values during our experiments. Another important point is that, in Step 3, the argument positions at which the given noun can appear is restricted to syntactic objects. This was because we empirically found that the rules generated from such verb-noun pairs were relatively accurate.

Assume that a given noun is “goods” and the verb pair “sell” and “manufacture” is selected in Step 3. Then, the following rule is generated.

- If someone *sells goods*, usually someone *manufactures* the *goods* at the same time as or before the event of the *selling* of the *goods*.

Although the word “someone” occurs twice, we do not demand that it refers to the same person in both instances. It just works as a placeholder. Also note that the adverb “usually”[1] was inserted to prevent the rule from being regarded as invalid by considering situations that are logically possible but unlikely in practice.

The above rule is produced when “manufacture” and “sell” frequently co-occur in coordinated sentences such as “The company *manufactured* goods and it *sold* them”. One might be puzzled because the order of the occurrences of the verbs in the coordinated sentences is reversed in the rule. The verb “sell” in the *second* (embedded) sentence/clause in the coordinated sentence appears as a verb in the precondition of the rule, while “manufacture” in the *first* (embedded) sentence/clause is the verb in the consequence.

A question then, is why we chose such an order, or such a direction of implication. There is another possibility, which might seem more straightforward. From the same coordinated sentences, we could produce the rule where the direction is reversed; i.e.,, “If someone *manufactures goods*, usually someone *sells*

[1] We used “`futsuu`” as a Japanese translation.

the *goods* at the same time as or *after* the manufacturing". The difference is that the rules generated by our procedure basically *infer* a *past* event from another event, while the rules with the opposite direction have to *predict* a future event. In experiments using our development set, we observed that the rules predicting future events were often unacceptable because of the uncertainty that we usually encounter in predicting the future or achieving a *future goal*. For instance, people might do something (e.g., manufacturing) with an intention to achieve some other goal (e.g., selling) in the future. But they sometimes fail to achieve their future goal for some reason. Some manufactured goods are never sold because, for instance, they are not good enough. In our experiments, we found that the precision rates of the rules with the direction we adopted were much higher than those of the rules with the opposite direction.

2.2 Simplified Scoring Function

To be precise, a rule generated by our method has the following form, where v_{pre} and v_{con} are verbs and n is a given noun.

- If someone v_{pre} n, usually someone v_{con} the n at the same time as or before the v_{pre}-ing of the n.

We assume that all three occurrences of noun n in the rule refer to the *same entity*.

Now, we define a simplified version of our scoring function as follows.

$$BasicS(n, v_{con}, v_{pre}, arg, arg') = \\ P_{coord}(v_{con}, v_{pre}) P_{arg'}(n|v_{pre}) P_{arg}(n|v_{con})/P(n)^2$$

Here, $P_{coord}(v_{con}, v_{pre})$ is the probability that v_{con} and v_{pre} are observed in coordinated sentences in a way that the event described by v_{con} temporally precedes or occurs at the same time as the event described by v_{pre}. (More precisely, v_{con} and v_{pre} must be the main verbs of two conjuncts S_1 and S_2 in a Japanese coordinated sentence that is literally translated to the form "S_1 *and* S_2".) This means that in the coordinated sentences, v_{con} appears first and v_{pre} second. $P_{arg'}(n|v_{pre})$ and $P_{arg}(n|v_{con})$ are the conditional probabilities that n occupies the argument positions arg' of v_{pre} and arg of v_{con}, respectively. At the beginning, as possible argument positions, we specified five argument positions, including the syntactic object and the subject. Note that when v_{pre} and v_{con} frequently co-occur in coordinated sentences and n often becomes arguments of v_{pre} and v_{con}, the score has a large value. This means that the score embodies our assumptions for acquiring rules.

The term $P_{coord}(v_{con}, v_{pre}) P_{arg'}(n|v_{pre}) P_{arg}(n|v_{con})$ in $BasicS$ is actually an approximation of the probability $P(v_{pre}, arg', n, v_{con}, arg, n)$ that we will observe the coordinated sentences such that the two sentences/clauses in the coordinated sentence are headed by v_{pre} and v_{con} and n occupies the argument positions arg' of v_{pre} and arg of v_{con}. Another important point is that the score is divided by $P(n)^2$. This is because the probabilities such as $P_{arg}(n|v_{con})$ tend to be large for a frequently observed noun n. The division by $P(n)^2$ is done to cancel such a tendency. This division does not affect the ranking for the same noun, but, since we give a *uniform* threshold for selecting the verb pairs for distinct nouns, such *normalization* is desirable, as we confirmed in experiments using our development set.

2.3 Paraphrases and Coordinated Sentences

Thus, we have defined our algorithm and a simplified scoring function. Now let us discuss a problem that is caused by the scoring function.

As mentioned in the introduction, a large portion of the acquired rules actually consists of *paraphrases*. Here, by a paraphrase, we mean a rule consisting of two descriptions referring to an identical event. The following example is an English translation of such paraphrases obtained by our method. We think this rule is acceptable. Note that we *invented* a new English verb "clearly-write" as a translation of a Japanese verb `meiki-suru` while "write" is a translation of another Japanese verb `kaku`.

- If someone clearly-writes a phone number, usually someone writes the phone number at the same time as or before the clearly-writing of the phone number.

Note that "clearly-write" and "write" have almost the same meaning but the former is often used in texts related to legal matters. Evidently, in the above rule, "clearly-write" and "write" describe the same event, and it can be seen as a *paraphrase*. There are two types of coordinated sentence that our method can use as clues to generate the rule.

- He clearly-wrote a *phone number* and wrote the *phone number*.
- He clearly-wrote a phone number, and also wrote an address.

The first sentence is more *similar* to the inference rule than the second in the sense that the two verbs share the same object. However, it is ridiculous because it describes the same event twice. Such a sentence is not observed frequently in corpora, and will not be used as clues to generate rules in practice.

On the other hand, we frequently observe sentences of the second type in corpora, and our method generates the paraphrases from the verb-verb co-occurrences taken from such sentences. However, there is a *mismatch* between the sentence and the acquired rule in the sense that the rule describes two events related to the same object (i.e., a phone number), while the above sentence describes two events that are related to distinct objects (i.e., a phone number and an address). Regarding this mismatch, two questions need to be addressed.

The first question is *why* our method *can* acquire the rule despite the mismatch. The answer is that our method obtains the verb-verb co-occurrence probabilities ($P_{coord}(v_{con}, v_{pre})$) and the verb-noun co-occurrence probabilities (e.g., $P_{arg}(n|v_{con})$) independently, and that the method does not check whether the two verbs share an argument.

Then the next question is why our method can acquire *accurate* paraphrases from such coordinated sentences. Though we do not have a definite answer now, our hypothesis is related to the strategy that people adopt in writing coordinated sentences. When two similar but distinct events, which *can* be described by the same verb, occur successively or at the same time, people avoid repeating the same verb to describe the two events in a single sentence. Instead they try to use distinct verbs that have similar meanings. Suppose that a person wrote his name and address. To report what she did, she may write "I clearly-wrote my name and also wrote my address" but will seldom write "I clearly-wrote my name and also clearly-wrote my address". Thus, we can expect to be able to find in coordinated sentences a large number of verb pairs consisting of two verbs with similar meanings. Note that our method tends to produce two verbs that frequently co-occur with a given noun. This also helps to produce the inference rules consisting of two semantically similar verbs.

3 Bias Mechanism

We now describe a bias used in our *full* scoring function, which significantly improves the precision. The full scoring function is defined as

$$Score(n, v_{con}, v_{pre}, arg, arg') = \\ P_{arg}(v_{con}) BasicS(n, v_{con}, v_{pre}, arg, arg').$$

The bias is denoted as $P_{arg}(v_{con})$, which is the probability that we can observe the verb v_{con}, which is the verb in the consequence of the rule, and its argument position arg is occupied by a noun, no matter which noun actually occupies the position.

An intuitive explanation of the assumption behind this bias is that as the situation within which the description of the consequence in a rule is valid becomes wider, the rule becomes more likely to be a proper one. Consider the following rules.

- If someone *demands* a compensation payment, someone *orders* the compensation payment.
- If someone *demands* a compensation payment, someone *requests* the compensation payment.

We consider the first rule to be unacceptable while the second expresses a proper implication. The difference is the situations in which the descriptions in the consequences hold. In our view, the situations described by "*order*" are more specific than those referred to by "*request*". In other words, "*order*" holds in a smaller range of situations than "*request*". *Requesting* something can happen in any situations where there exists someone who can demand something, but *ordering* can occur only in a situations where someone in a particular social position can demand something. The basic assumption behind our bias is that rules with consequences that can be valid in a *wider* range of situations, such as "requesting a compensation payment," are more likely to be proper ones than the rules with consequences that hold in a *smaller* range of situations, such as "ordering a compensation payment".

The bias $P_{arg}(v_{con})$ was introduced to capture variations of the situations in which event descriptions are valid. We assume that frequently observed verbs form *generic* descriptions that can be valid within a wide range of events, while less frequent verbs tend to describe events that can occur in a narrower range of situations and form more specific descriptions than the frequently observed verbs. Regarding the "request-order" example, (a Japanese translation of) "request" is observed more frequently than (a Japanese translation of) "order" in corpora and this observation is consistent with our assumption. A similar idea by Geffet and Dagan (Geffet and Dagan, 2005) was proposed for capturing lexical entailment. The difference is that they relied on word co-occurrences rather than the frequency of words to measure the specificity of the semantic contents of lexical descriptions, and needed Web search to avoid data sparseness in co-occurrence statistics. On the other hand, our method needs only simple occurrence probabilities of single verbs and we expect our method to be applicable to wider vocabulary than Geffet and Dagan's method.

The following is a more mathematical justification for the bias. According to the following discussion, $P_{arg}(v_{con})$ can be seen as a metric indicating how *easily* we can *establish* an interpretation of the rule, which is formalized as a mapping between events. In our view, if we can establish the mapping *easily*, the rule tends to be acceptable. The discussion starts from a formalization of an *interpretation* of an inference rule. Consider the rule "If exp_1 occurs, usually exp_2 occurs at the same time or before the occurrence of exp_1", where exp_1 and exp_2 are natural language expressions referring to events. In the following, we call such expressions *event descriptions* and distinguish them from an *actual event* referred to by the expressions. An actual event is called an *event instance*.

A possible interpretation of the rule is that, for any event instance e_1 that can be described by the event description exp_1 in the precondition of the rule, there always exists an event instance e_2 that can be described by the event description exp_2 in the consequence and that occurs at the same time as or before e_1 occurs. Let us write $e : exp$ if event instance e can be described by event description exp. The above interpretation can then be represented by the formula

$$\Phi : \exists f(\forall e_1(e_1 : exp_1 \rightarrow \exists e_2(e_2 = f(e_1) \wedge e_2 : exp_2)).$$

Here, the mapping f represents a temporal relation between events, and the formula $e_2 = f(e_1)$ expresses that e_2 occurs at the same time as or before e_1.

The bias $P_{arg}(v_{con})$ can be considered (an approximation of) a parameter required for computing the probability that a mapping f_{random} satisfies the requirements for f in Φ when we randomly *construct* f_{random}. The probability is denoted as $P\{e_2 : exp_2 \wedge e_2 = f_{random}(e_1)|e_1 : exp_1\}^{E_1}$ where E_1 denotes the number of events describable by exp_1. We assume that the larger this probability is, the more easily we can establish f. We can approximate $P\{e_2 : exp_2 \wedge e_2 = f_{random}(e_1)|e_1 : exp_1\}$ as $P(exp_2)$ by 1) observing that the probabilistic variables e_1 and e_2 are independent since f_{random} associates them in a completely random manner and by 2) assuming that the occurrence probability of the event instances describable by exp_2 can be approximated by the probability that exp_2 is observed in text corpora. This means that $P(exp_2)$ is one of the metrics indicating how *easily* we can establish the mapping f in Φ.

Then, the next question is what kind of expressions should be regarded as the event description exp_2. A primary candidate will be the whole sentence appearing in the consequence part of the rule to be produced. Since we specify only a verb v_{con} and its argument n in the consequence in a rule, $P(exp_2)$ can be denoted by $P_{arg}(n, v_{con})$, which is the probability that we observe the expression such that v_{con} is a head verb and n occupies an argument position arg of v_{con}. By multiplying this probability to $BasicS$ as a bias, we obtain the following scoring function.

$$Score_{cooc}(n, v_{con}, v_{pre}, arg, arg') = \\ P_{arg}(n, v_{con})BasicS(n, v_{con}, v_{pre}, arg, arg')$$

In our experiments, though, this score did not work well. Since $P_{arg}(n, v_{con})$ often has a small value, the problem of data sparseness seems to arise. Then, we used $P_{arg}(v_{con})$, which denotes the probability of observing sentences that contain v_{con} and its argument position arg, no matter which noun occupies arg, instead of $P_{arg}(n, v_{con})$. We multiplied the probability to $BasicS$ as a bias and obtained the following score, which is actually the scoring function we propose.

$$Score(n, v_{con}, v_{pre}, arg, arg') = \\ P_{arg}(v_{con})BasicS(n, v_{con}, v_{pre}, arg, arg')$$

4 Experiments

4.1 Settings

We parsed 35 years of newspaper articles (Yomiuri 87-01, Mainichi 91-99, Nikkei 90-00, 3.24GB in total) and 92.6GB of HTML documents downloaded from the WWW using an existing parser (Kanayama et al., 2000) to obtain the word (co-occurrence) frequencies. All the probabilities used in our method were estimated by maximum likelihood estimation from these frequencies. We randomly picked 600 nouns as a development set. We prepared three test sets, namely test sets A, B, and C, which consisted of 100 nouns, 250 nouns and 1,000 nouns respectively. Note that all the nouns in the test sets were randomly picked and did not have any common items with the development set. In all the experiments, four human judges checked if each produced rule was a proper one without knowing how each rule was produced.

4.2 Effects of Using Coordinated Sentences

In the first series of experiments, we compared a simplified version of our scoring function $BasicS$ with some alternative scores. This was mainly to check if coordinated sentences can improve accuracy. The alternative scores we considered

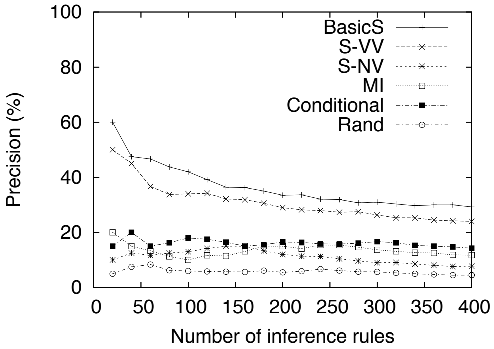

Figure 1: Comparison with the alternatives (4 judges)

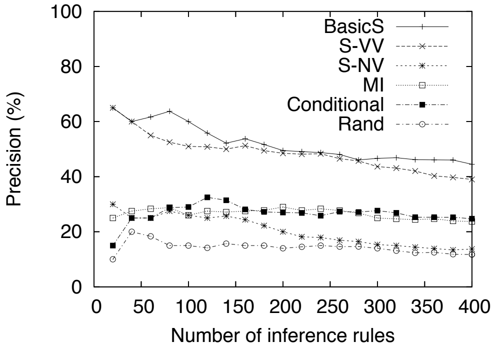

Figure 2: Comparison with the alternatives (3 judges)

are presented below. Note that we did not test our bias mechanism in this series of experiments.

$$S\text{-}VV(n, v_{con}, v_{pre}, arg, arg') = P_{arg}(n, v_{con}) P_{arg'}(n, v_{pre}) / P(n)^2$$
$$S\text{-}NV(n, v_{con}, v_{pre}) = P_{coord}(v_{con}, v_{pre})$$
$$MI(n, v_{con}, v_{pre}) = P_{coord}(v_{con}, v_{pre}) / (P(v_{con}) P(v_{pre}))$$
$$Cond(n, v_{con}, v_{pre}, arg, arg') = P_{coord}(v_{con}, v_{pre}, arg, arg') P_{arg}(n|v_{con}) P_{arg'}(n|v_{pre}) / (P_{arg'}(n, v_{pre}) P(n))$$
$$Rand(n, v_{con}, v_{pre}, arg, arg') = \text{random number}$$

$S\text{-}VV$ was obtained by approximating the probabilities of coordinated sentences, as in the case of $BasicS$. However, we assumed the occurrences of two verbs were independent. The difference between the performance of this score and that of $BasicS$ will indicate the effectiveness of using verb-verb co-occurrences in coordinated sentences.

The second alternative, $S\text{-}NV$, simply ignores the noun-verb co-occurrences in $BasicS$. MI is a score based on mutual information and roughly corresponds to the score used in a previous attempt to acquire temporal relations between events (Chklovski and Pantel, 2004). $Cond$ is an approximation of the probability $P(n, v_{con}|n, v_{pre})$; i.e., the conditional probability that the coordinated sentences consisting of n, v_{con} and v_{pre} are observed given the precondition part consisting of v_{pre} and n. $Rand$ is a random number and generates rules by combining verbs that co-occur with the given n randomly. This was used as a baseline method of our task

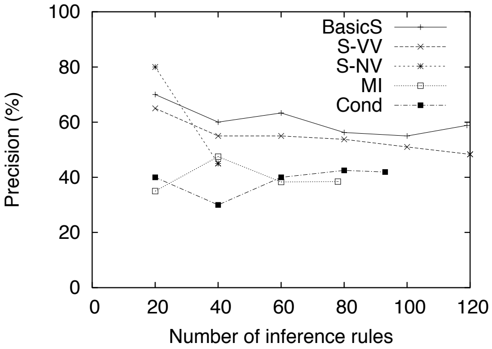

Figure 3: Comparison with the alternatives (3 judges)

The resulting precisions are shown in Figures 1 and 2. The figure captions specify "(4 judges)", as in Figure 1, when the acceptable rules included only those regarded as proper by all four judges; the captions specify "(3 judges)", as in Figure 2, when the acceptable rules include those considered proper by at least three of the four judges. We used test set A (100 nouns) and produced the top four rule candidates for each noun according to each score. As the final results, all the produced rules for all the nouns were sorted according to each score, and a precision was obtained for top N rules in the sorted list. This was the same as the precision achieved by setting the score value of N-th rule in the sorted list as threshold θ. Notice that $BasicS$ outperformed all the alternatives[2], though the difference between $S\text{-}VV$ and $BasicS$ was rather small. Another important point is that the precisions obtained with the scores that ignored noun-verb co-occurrences were quite low. These findings suggest that 1) coordinated sentences can be useful clues for obtaining temporally constrained rules and 2) noun-verb co-occurrences are also important clues.

In the above experiments, we actually allowed noun n to appear as argument types other than the syntactic objects of a verb. When we restricted the argu-

[2] Actually, the experiments concerning $Rand$ were conducted considerably after the experiments on the other scores, and only the two of the four judges for $Rand$ were included in the judges for other scores. However, we think that the superiority of our score $BasicS$ over the baseline method was confirmed since the precision of $Rand$ was drastically lower than that of $BasicS$

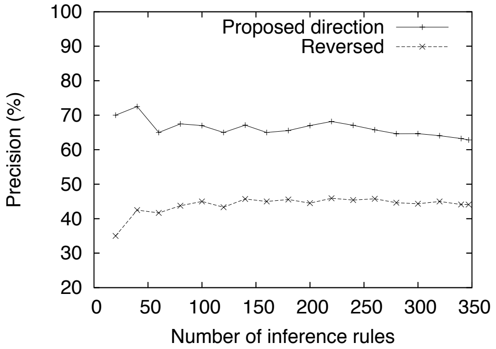

Figure 4: Two directions of implications (3 judges)

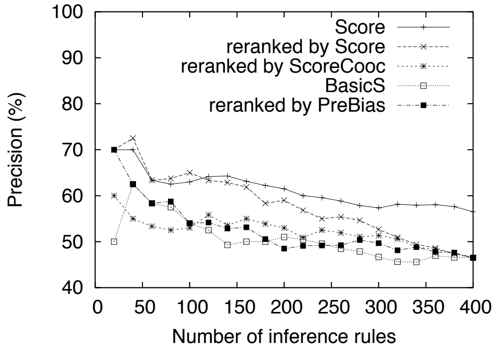

Figure 5: Effects of the bias (4 judges)

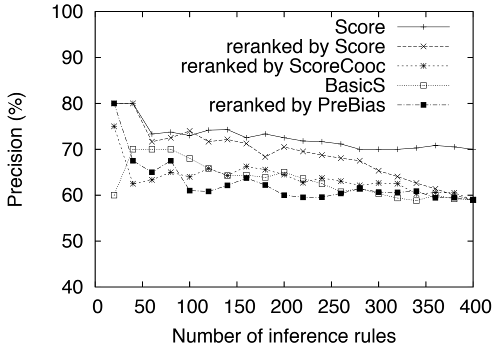

Figure 6: Effects of the bias (3 judges)

ment types to syntactic objects, as described in Section 2, the precision shown in Figure 3 was obtained. In most cases, $BasicS$ outperformed the alternatives. Although the number of produced rules was reduced because of this restriction, the precision of all produced rules was improved. Because of this, we decided to restrict the argument type to objects.

The kappa statistic for assessing the inter-rater agreement was 0.53, which indicates *moderate* agreement according to Landis and Koch, 1977. The kappa value for only the judgments on rules produced by $BasicS$ rose to 0.59. After we restricted the verb-noun co-occurrences to verb-object co-occurrences, the kappa became 0.49, while that for the rules produced by $BasicS$ was 0.54[3].

4.3 Direction of Implications

Next, we examined the directions of implications and the temporal order between events. We produced 1,000 rules for test set B (250 nouns) using the score $BasicS$, again without restricting the argument types of given nouns to syntactic objects. When we restricted the argument positions to objects, we obtained 347 rules. Then, from each generated rule, we created a new rule having an opposite direction of implications. We swapped the precondition and the consequence of the rule and reversed its temporal order. For instance, we created "If someone enacts a law, usually someone enforces the law at the same time as or *after* the enacting of the law" from "If someone enforces a law, usually someone enacts the law at the same time as or before the enforcing of the law".

Figure 4 shows the results. 'Proposed direction' refers to the precision of the rules generated by our method. The precision of the rules with the opposite direction is indicated by 'Reversed.' The precision of 'Reversed' was much lower than that of our method, and this justifies our choice of direction. The kappas values for 'BasicS' and 'Reversed' were 0.54 and 0.46 respectively. Both indicate moderate agreement.

4.4 Effects of the Bias

Last, we compared $Score$ and $BasicS$ to see the effect of our bias. This time, we used test set C (1,000 nouns). The rules were restricted to those in which the given nouns are syntactic objects of two verbs. The evaluation was done for only the top 400 rules for each score. The results are shown in Figures 5 and 6. 'Score' refers to the precision obtained with $Score$, while 'BasicS' indicates the precision with $BasicS$. For most data points in both graphs, the 'Score' precision was about 10% higher than the 'BasicS' precision. In Figure 6, the precision reached 70% when the 400 rules were produced. These results indicate the desirable effect of our bias for, at least, the top rules.

[3] These kappa values were calculated for the results except for the ones obtained by the score $Rand$, which were assessed by different judges. The kappa for $Rand$ was 0.33 (fair agreement).

rank /judges	inference rules
4/0	`moshi yougi wo hininsuru naraba, yougi wo mitomeru` (If someone denies suspicions, usually someone confirms the suspicions.)
6/4	`moshi jikokiroku wo uwamawaru naraba, jikokiroku wo koushinsuru` (If someone betters her best record, usually someone breaks her best record.)
21/3	`moshi katakuriko wo mabusu naraba, katakuriko wo tsukeru` (If someone coats something with potato starch, usually someone covers something with the starch)
194/4	`moshi sasshi wo haifusuru naraba, sasshi wo sakuseisuru` (If someone distributes a booklet, usually someone makes the booklet.)
303/4	`moshi netsuzou wo kokuhakusuru naraba, netsuzou wo mitomeru` (If someone confesses to a fabrication, usually someone admits the fabrication.)
398/3	`moshi ifuku wo kikaeru naraba, ifuku wo nugu` (If someone changes clothes, usually someone gets out of the clothes.)

Figure 7: Examples of acquired inference rules

The 400 rules generated by $Score$ included 175 distinct nouns and 272 distinct verb pairs. Examples of the inference rules acquired by $Score$ are shown in Figure 7 along with the positions in the ranking and the numbers of judges who judged the rule as being proper. (We omitted the phrase "the same time as or before" in the examples.) The kappa was 0.57 (moderate agreement).

In addition, the graphs compare $Score$ with some other alternatives. This comparison was made to check the effectiveness of our bias more carefully. The 400 rules generated by $BasicS$ were re-ranked using $Score$ and the alternative scores, and the precision for each was computed using the human judgments for the rules generated by $BasicS$. (We did not evaluate the rules directly generated by the alternatives to reduce the workload of the judges.) The first alternative was $Score_{cooc}$, which was presented in Section 3. Here, "reranked by ScoreCooc" refers to the precision obtained by re-ranking with of $Score_{cooc}$. The precision was below that obtained by the re-ranking with $Score$, (referred to as "reranked by Score)". As discussed in Section 3, this indicates the bias $P_{arg}(v_{con})$ in $Score$ works better than the bias $P_{arg}(n, v_{con})$ in $Score_{cooc}$.

The second alternative was the scoring function obtained by replacing the bias $P_{arg}(v_{con})$ in $Score$ with $P_{arg'}(v_{pre})$, which is roughly the probability that the verb in the precondition will be observed. The score is denoted as $PreBias(n, v_{con}, v_{pre}, arg, arg') = P_{arg'}(v_{pre})BasicS(n, v_{con}, v_{pre}, arg, arg')$. The precision of this score is indicated by "reranked by PreBias" and is much lower than that of "reranked by Score", indicating that only probability of the verbs in the consequences should be used as a bias. This is consistent with our assumption behind the bias.

5 Conclusion

We have presented an unsupervised method for acquiring inference rules with temporal constraints, such as "If someone enforces a law, someone enacts the law at the same time as or before the enforcing of the law". We used the probabilities of verb-verb co-occurrences in coordinated sentences and verb-noun co-occurrences. We have also proposed a bias mechanism that can improve the precision of acquired rules.

References

R. Barzilay and L. Lee. 2003. Learning to paraphrase:an unsupervised approach using multiple-sequence alignment. In *Proc. of HLT-NAACL 2003*, pages 16–23.

T. Chklovski and P. Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proc. of EMNLP-04*.

I. Dagan, O. Glickman, and B. Magnini, editors. 2005. *Proceedings of the First Challenge Workshop: Recognizing Textual Entailment*. available from http://www.pascal-network.org/Challenges/RTE/.

T. Fujiki, H. Namba, and M. Okumura. 2003. Automatic acquisition of script knowledge from text collection. In *Proc. of The Research Note Sessions of EACL'03*.

M. Geffet and I. Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proc. of ACL 2005*, pages 107–114.

H. Kanayama, K. Torisawa, Y. Mitsuishi, and J. Tsujii. 2000. A hybrid Japanese parser with hand-crafted grammar and statistics. In *Proc. of COLING 2000*.

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

D. Lin and P. Pantel. 2001. Discovery of inference rules for question answering. *Journal of Natural Language Engineering*.

Y. Shinyama, S. Sekine, and K. Sudo. 2002. Automatic paraphrase acquisition from news articles. In *Proc. of HLT2002*.

I. Szepektor, H. Tanev, I. Dagan, and B. Coppola. 2004. Scaling web-based acquisition of entailment relations. In *Proc. of EMNLP 2004*.