Large-Scale Verb Entailment Acquisition from the Web

Chikara Hashimoto* **Kentaro Torisawa**† **Kow Kuroda**‡
Stijn De Saeger§ **Masaki Murata**¶ **Jun'ichi Kazama**$^\|$
National Institute of Information and Communications Technology
Sorakugun, Kyoto, 619-0289, JAPAN
{*ch,†torisawa,‡kuroda,§stijn,¶murata,$^\|$kazama}@nict.go.jp

Abstract

Textual entailment recognition plays a fundamental role in tasks that require in-depth natural language understanding. In order to use entailment recognition technologies for real-world applications, a large-scale entailment knowledge base is indispensable. This paper proposes a conditional probability based directional similarity measure to acquire verb entailment pairs on a large scale. We targeted 52,562 verb types that were derived from 10^8 Japanese Web documents, without regard for whether they were used in daily life or only in specific fields. In an evaluation of the top 20,000 verb entailment pairs acquired by previous methods and ours, we found that our similarity measure outperformed the previous ones. Our method also worked well for the top 100,000 results.

1 Introduction

We all know that if you snored, you must have been sleeping, that if you are divorced, you must have been married, and that if you won a lawsuit, you must have sued somebody. These relationships between events where one is the logical consequence of the other are called entailment. Such knowledge plays a fundamental role in tasks that require in-depth natural language understanding, e.g., answering questions and using natural language interfaces.

This paper proposes a novel method for verb entailment acquisition. Using a Japanese Web corpus (Kawahara and Kurohashi, 2006a) derived from 10^8 Japanese Web documents, we automatically acquired such verb pairs as *snore* → *sleep* and *divorce* → *marry*, where entailment holds between the verbs in the pair.[1] Our definition of "entailment" is the same as that in WordNet3.0; v_1 entails v_2 if v_1 cannot be done unless v_2 is, or has been, done.[2]

Our method follows the distributional similarity hypothesis, i.e., words that occur in the same context tend to have similar meanings. Just as in the methods of Lin and Pantel (2001) and Szpektor and Dagan (2008), we regard the arguments of verbs as the context in the hypothesis. However, unlike the previous methods, ours is based on conditional probability and is augmented with a simple trick that improves the accuracy of verb entailment acquisition. In an evaluation of the top 20,000 verb entailment pairs acquired by the previous methods and ours, we found that our similarity measure outperformed the previous ones. Our method also worked well for the top 100,000 results,

Since the scope of Natural Language Processing (NLP) has advanced from a formal writing style to a colloquial style and from restricted to open domains, it is necessary for the language resources for NLP, including verb entailment knowledge bases, to cover a broad range of expressions, regardless of whether they are used in daily life or only in specific fields that are highly technical. As we will discuss later, our method can acquire, with reasonable accuracy, verb entailment pairs that deal not only with common and familiar verbs but also with technical and unfamiliar ones like *podcast* → *download* and *jibe* → *sail*.

Note that previous researches on entailment acquisition focused on templates with variables or word-lattices (Lin and Pantel, 2001; Szpektor and Dagan, 2008; Barzilay and Lee, 2003; Shinyama

[1] Verb entailment pairs are described as $v_1 \rightarrow v_2$ (v_1 is the entailing verb and v_2 is the entailed one) henceforth.

[2] WordNet3.0 provides entailment relationships between synsets like *divorce, split up* → *marry, get married, wed, conjoin, hook up with, get hitched with, espouse.*

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1172–1181,
Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

et al., 2002). Certainly these templates or word lattices are more useful in such NLP applications as Q&A than simple entailment relations between verbs. However, our contention is that entailment certainly holds for some verb pairs (like *snore* → *sleep*) by themselves, and that such pairs constitute the core of a future entailment rule database. Although we focused on verb entailment, our method can also acquire template-level entailment pairs with a reasonable accuracy.

The rest of this paper is organized as follows. In §2, related works are described. §3 presents our proposed method. After this, an evaluation of our method and the existing methods is presented in Section 4. Finally, we conclude the paper in §5.

2 Related Work

Previous studies on entailment, inference rules, and paraphrase acquisition are roughly classified into those that require comparable corpora (Shinyama et al., 2002; Barzilay and Lee, 2003; Ibrahim et al., 2003) and those that do not (Lin and Pantel, 2001; Weeds and Weir, 2003; Geffet and Dagan, 2005; Pekar, 2006; Bhagat et al., 2007; Szpektor and Dagan, 2008).

Shinyama et al. (2002) regarded newspaper articles that describe the same event as a pool of paraphrases, and acquired them by exploiting named entity recognition. They assumed that named entities are preserved across paraphrases, and that text fragments in the articles that share several comparable named entities should be paraphrases. Barzilay and Lee (2003) also used newspaper articles on the same event as comparable corpora to acquire paraphrases. They induced paraphrasing patterns by sentence clustering. Ibrahim et al. (2003) relied on multiple English translations of foreign novels and sentence alignment to acquire paraphrases. We decided not to take this approach since using comparable corpora limits the scale of the acquired paraphrases or entailment knowledge bases. Although obtaining comparable corpora has been simplified by the recent explosion of the Web, the availability of plain texts is incomparably better.

Entailment acquisition methods that do not require comparable corpora are mostly based on the distributional similarity hypothesis and use plain texts with a syntactic parser. Basically, they parse texts to obtain pairs of predicate phrases and their arguments, which are regarded as features of the predicates with appropriately assigned weights. Lin and Pantel (2001) proposed a paraphrase acquisition method (non-directional similarity measure) called DIRT which acquires pairs of binary-templates (predicate phrases with two argument slots) that are paraphrases of each other. DIRT employs the following similarity measure proposed by Lin (1998):

$$Lin(l,r) = \frac{\sum_{f \in F_l \cap F_r} [w_l(f) + w_r(f)]}{\sum_{f \in F_l} w_l(f) + \sum_{f \in F_r} w_r(f)}$$

where l and r are the corresponding slots of two binary templates, F_s is s's feature vector (argument nouns), and $w_s(f)$ is the weight of $f \in F_s$ (PMI between s and f). The intuition behind this is that the more nouns two templates share, the more semantically similar they are. Since we acquire verb entailment pairs based on unary templates (Szpektor and Dagan, 2008) we used the Lin formula to acquire unary templates directly rather than using the DIRT formula, which is the arithmetic-geometric mean of Lin's similarities for two slots in a binary template.

Bhagat et al. (2007) developed an algorithm called LEDIR for learning the directionality of non-directional inference rules like those produced by DIRT. LEDIR implements a Directionality Hypothesis: when two binary semantic relations tend to occur in similar contexts and the first one occurs in significantly more contexts than the second, then the second most likely implies the first and not vice versa.

Weeds and Weir (2003) proposed a general framework for distributional similarity that mainly consists of the notions of what they call Precision (defined below) and Recall:

$$Precision(l,r) = \frac{\sum_{f \in F_l \cap F_r} w_l(f)}{\sum_{f \in F_l} w_l(f)}$$

where l and r are the targets of a similarity measurement, F_s is s's feature vector, and $w_s(f)$ is the weight of $f \in F_s$. The best performing weight is PMI. Precision is a directional similarity measure that examines the coverage of l's features by those of r's, with more coverage indicating more similarity.

Szpektor and Dagan (2008) proposed a directional similarity measure called BInc (Balanced-Inclusion) that consists of Lin and Precision, as

$$BInc(l,r) = \sqrt{Lin(l,r) \times Precision(l,r)}$$

where l and r are the target templates. For weighting features, they used PMI. Szpektor and Dagan (2008) also proposed a unary template, which is defined as a template consisting of one argument slot and one predicate phrase. For example, *X take a nap* → *X sleep* is an entailment pair consisting of two unary templates. Note that the slot *X* must be shared between templates. Though most of the previous entailment acquisition studies focused on binary templates, unary templates have an obvious advantage over binary ones; they can handle intransitive predicate phrases and those that have omitted arguments. The Japanese language, which we deal with here, often omits arguments, and thus the advantage of unary templates is obvious.

As shown in §4, our method outperforms Lin, Precision, and BInc in accuracy.

Szpector et al. (2004) addressed broad coverage entailment acquisition. But their method requires an existing lexicon to start, while ours does not.

Apart from the dichotomy of the comparable corpora and the distributional similarity approaches, Torisawa (2006) exploited the structure of Japanese coordinated sentences to acquire verb entailment pairs. Pekar (2006) used the local structure of coherent text by identifying related clauses within a local discourse. Zanzotto et al. (2006) exploited agentive nouns. For example, they acquired *win* → *play* from "*the player wins*."

Geffet and Dagan (2005) proposed the Distributional Inclusion Hypotheses, which claimed that if a word *v* entails another word *w*, then all the characteristic features of *v* are expected to appear with *w*, and vice versa. They applied this to noun entailment pair acquisition, rather than verb pairs.

3 Proposed Method

This section presents our method of verb entailment acquisition. First, the basics of Japanese are described. Then, we present the directional similarity measure that we developed in §3.2. §3.3 describes the structure and acquisition of the web-based data from which entailment pairs are derived. Finally, we show how we acquire verb entailment pairs using our proposed similarity measure and the web-based data in §3.4.

3.1 Basics of Japanese

Japanese explicitly marks arguments including the subject and object by postpositions, and is a head-final language. Thus, a verb phrase consisting of an object *hon* (book) and a verb *yomu* (read), for example, is expressed as *hon-wo yomu* (book-ACC read) "read a book" with the accusative postposition *wo* marking the object.[3] Accordingly, we refer to a unary template as $\langle p, v\rangle$ hereafter, with p and v referring to the *p*ostposition and a *v*erb. Also, we abbreviate a template-level entailment $\langle p_l, v_l\rangle \rightarrow \langle p_r, v_r\rangle$ as $l \rightarrow r$ for simplicity. We define a unary template as a template consisting of one argument slot and one predicate, following Szpektor and Dagan (2008).

3.2 Directional Similarity Measure based on Conditional Probability

The directional similarity measure that we developed and called $Score$ is defined as follows:

$$Score(l, r) = Score_{base}(l, r) \times Score_{trick}(l, r)$$

where l and r are unary templates, and $Score$ indicates the probability of $l \rightarrow r$. $Score_{base}$, which is the base of $Score$, is defined as follows:

$$Score_{base}(l, r) = \sum_{f \in F_l \cap F_r} P(r|f)P(f|l)$$

where F_s is s's feature vector (nouns including compounds). The intention behind the definition of $Score_{base}$ is to *emulate* the conditional probability $P(v_r|v_l)$[4] in a distributional similarity style function. Note that $P(v_r|v_l)$ should be 1 when entailment $v_l \rightarrow v_r$ holds (i.e., v_r is observed whenever v_l is observed) and we have reliable probability values. Then, if we can directly estimate $P(v_r|v_l)$, it is reasonable to assume $v_l \rightarrow v_r$ if $P(v_r|v_l)$ is large enough. However, we cannot estimate $P(v_r|v_l)$ directly since it is unlikely that we will observe the verbs v_r and v_l at the same time. (People do not usually repeat v_r and v_l in the same document to avoid redundancy.) Thus, instead of a direct estimation, we *substitute* $Score_{base}(l, r)$ as defined above. In other words, we assume $P(v_r|v_l) \approx P(r|l) \approx \Sigma_{f \in F_l \cap F_r} P(f|l)P(r|f)$.

Actually, $Score_{base}$ originally had another motivation, inspired by Torisawa (2005), for which no postposition but the instrumental postposition *de* was relevant. In this discussion, all of the nouns (fs) that are marked by the instrumental postposition are seen as "tools," and $P(f|l)$ is interpreted

[3] ACC represents an accusative postposition in Japanese. Likewise, NOM, DAT, INS, and TOP are the symbols for the nominative, dative, instrumental, and topic postpositions.

[4] Remember that v_l and v_r are the verbs of unary templates l and r.

as a measure of how typically the tool f is used to perform the action denoted by (the v_l of) l; if $P(f|l)$ is large enough, f is a typical tool used in l. On the other hand, $P(r|f)$ indicates the probability of (the v_r of) r being the purpose for using the tool f. See (1) for an example.

(1) *konro-de* *chouri-suru*
cooking.stove-INS cook
'cook (something) using a cooking stove.'

The purpose of using a *cooking stove* is to *cook*. Torisawa (2005) has pointed out that when r expresses the purpose of using a tool f, $P(r|f)$ tends to be large. This predicts that $P(r|\textit{cooking stove})$ is large, where r is $\langle de, cook\rangle$.

According to this observation, if f is a single purpose tool and $P(f|l)$, the probability of f being the tool by which l is performed, and $P(r|f)$, the probability of r being the purpose of using the tool f, are large enough, then the typical performance of the action v_l should contain some actions that can be described by v_r, i.e., the purpose of using f. Moreover, if all the typical tools (fs) used in v_l are also used for v_r, most performances of the action v_l should contain a part described by the action v_r. In summary, this means that when $\Sigma_{f\in F_l\cap F_r}P(r|f)P(f|l)$, $Score_{base}$, has a large value, we can expect $v_l \rightarrow v_r$.

For example, let v_l be *deep-fry* and v_r be *cook*. Note that $v_l \rightarrow v_r$ holds for this example. There are many tools that are used for deep-frying, such as *cooking stove*, *pot*, or *pan*. This means that $P(\textit{cooking stove}|l)$, $P(\textit{pot}|l)$, or $P(\textit{pan}|l)$ are large. On the other hand, the purpose of using all of these tools is *cooking*, based on common sense. Thus, probabilities such as $P(r|\textit{cooking stove})$ and $P(r|\textit{pan})$ should have large values. Accordingly, $\Sigma_{f\in F_l\cap F_r}P(f|l)P(r|f)$, $Score_{base}$, should be relatively large for *deep-fry* → *cook,*

Actually, we defined $Score_{base}$ based on the above assumption However, through a series of preliminary experiments, we found that the same score could be applied without losing the precision to the other postpositions. Thus, we generalized the framework so that it could deal with most postpositions, namely *ga* (NOM), *wo* (ACC), *ni* (DAT), *de* (INS), and *wa* (TOP). Note that this is a variation of the distributional inclusion hypothesis (Geffet and Dagan, 2005), but that we do not use mutual information as in previous works, based on the hypothesis discussed above. Actually, as shown in §4, our conditional probability based method outperformed the mutual information based metrics in our experiments.

On the other hand, $Score_{trick}$ implements another assumption that if only one feature contributes to $Score_{base}$ and the contribution of the other nouns is negligible, if any, the similarity is unreliable. Accordingly, for $Score_{trick}$, we uniformly ignore the contribution of the most dominant feature from the similarity measurement.

$$\begin{aligned} & Score_{trick}(l,r) \\ & \quad = \; Score_{base}(l,r) - \max_{f\in F_l\cap F_r} P(r|f)P(f|l) \end{aligned}$$

As shown in §4, this trick actually improved the entailment acquisition accuracy.

We used maximum likelihood estimation to obtain $P(r|f)$ and $P(f|l)$ in the above discussion.

Bannard and Callison-Burch (2005) and Fujita and Sato (2008) also proposed directional similarity measures based on conditional probability, which are very similar to $Score_{base}$, although either their method's prerequisites or the targets of the similarity measurements were different from ours. The method of Bannard and Callison-Burch (2005) requires bilingual parallel corpora, and uses the translations of expressions as its feature. Fujita and Sato (2008) dealt with productive predicate phrases, while our target is non-productive lexical units, i.e., verbs. Thus, this is the first attempt to apply a conditional probability based similarity measure to verb entailment acquisition. In addition, the trick implemented in $Score_{trick}$ is novel.

3.3 Preparing Template-Feature Tuples

Our method starts from a dataset called template-feature tuples, which was derived from the Web in the following way: **1)** Parse the Japanese Web corpus (Kawahara and Kurohashi, 2006a) derived from 10^8 Japanese Web documents with Japanese dependency parser KNP (Kawahara and Kurohashi, 2006b). **2)** Extract triples $\langle n,p,v\rangle$ consisting of nouns (n), postpositions (p), and verbs (v), where an n marked by a p depends on a v from the parsed Web text. **3)** From the triple database, construct template-feature tuples $\langle n,\langle p,v\rangle\rangle$ by regarding $\langle p,v\rangle$ as a unary template and n as one of its features. **4)** Convert the verbs into their canonical forms as defined by KNP. **5)** Filter out tuples that fall into one of the following categories: 5-1) $Freq(\langle p,v\rangle) < 20$. 5-2) Its verb is passivized, causativized, or negated. 5-3) Its verb is semantically vague like *be*, *do*, or *become*. 5-4) Its postposition is something other than *ga* (NOM), *wo* (ACC), *ni* (DAT), *de* (INS), or *wa* (TOP).

The resulting unary template-feature tuples included 127,808 kinds of templates that consisted of 52,562 verb types and five kinds of postpositions. The verbs included compound words like *bosi-kansen-suru* (mother.to.child-infection-do) "infect from mothers to infants."

3.4 Acquiring Entailment Pairs

We acquired verb entailment pairs using the following procedure: **i)** From the template-feature tuples mentioned in §3.3, acquire unary template pairs that exhibit an entailment relation between them using the directional similarity measure in §3.2. **ii)** Convert the acquired unary templates $\langle p, v \rangle$ into naked verbs v by stripping the postpositions p. **iii)** Remove the duplicated verb pairs resulting from stripping ps. To be precise, when we removed the duplicated pairs, we left the highest ranked one. **iv)** Retrieve N-best verb pairs as the final output from the result of iii). That is, we first acquired unary template pairs and then transformed them into verb pairs.

Although this paper focuses on verb entailment acquisition, we also evaluated the accuracy of template-level entailment acquisition, in order to show that our similarity measure works well, not only for verb entailment acquisition, but also for template entailment acquisition (See §4.4). we created two kinds of unary templates: the "Scoring Slots" template and the "Nom(inative) Slots" template. The first is simply the result of the procedure **i)**; all of the templates have slots that are used for similarity scoring. The second one was obtained in the following way: **1)** Only templates whose p is not a nominative are sampled from the result of the procedure **i)**. **2)** Their ps are all changed to a nominative. Templates of the second kind are used to show that the corresponding slots between templates (nominative, in this case) that are not used for similarity scoring can be incorporated to resulting template-level entailment pairs if the scoring function really captures the semantic similarity between templates.

Note that, for unary template entailment pairs like (2) to be well-formed, the two unary slots (*X-wo*) between templates must share the same noun as the index i indicates. This is relevant in §4.4.

(2) *X_i-wo musaborikuu* → *X_i-wo taberu*
X_i-ACC gobble X_i-ACC eat

4 Evaluation

We compare the accuracy of our method with that of the alternative methods in §4.1. §4.2 shows the effectiveness of the trick. We examine the entailment acquisition accuracy for frequent verbs in §4.3, and evaluate the performance of our method when applied to template-level entailment acquisition in §4.4. Finally, by showing the accuracy for verb pairs obtained from the top 100,000 results, we claim that our method provides a good starting point from which a large-scale verb entailment resource can be constructed in §4.5.

For the evaluation, three human annotators (not the authors) checked whether each acquired entailment pair was correct. The average of the three Kappa values for each annotator pair was 0.579 for verb entailment pairs and 0.568 for template entailment pairs, both of which indicate the middling stability of this evaluation annotation.

4.1 Experiment 1: Verb Pairs

We applied $Score$, BInc, Lin, and Precision to the template-feature tuples (§3.3), obtained template entailment pairs, and finally obtained verb entailment pairs by removing the postpositions from the templates as described in §3. As a baseline, we created pairs from randomly chosen verbs.

Since we targeted all of the verbs that appeared on the Web (under the condition of $Freq(\langle p, v \rangle) \geq 20$), the annotators were confronted with technical terms and slang that they did not know. In such cases, they consulted dictionaries (either printed or machine readable ones) and the Web. If they still could not find the meaning of a verb, they labeled the pair containing the unknown verb as incorrect.

We used the accuracy $= \frac{\text{\# of correct pairs}}{\text{\# of acquired pairs}}$ as an evaluation measure. We regarded a pair as correct if it was judged correct by one (Accuracy-1), two (Accuracy-2), or three (Accuracy-3) annotators.

We evaluated 200 entailment pairs sampled from the top 20,000 for each method (# of acquired pairs = 200). For fairness, the evaluation samples for each method were shuffled and placed in one file from which the annotators worked. In this way, they were unable to know which entailment pair came from which method.

Note that the verb entailment pairs produced by Lin do not provide the directionality of entailment. Thus, the annotators decided the directionality of these entailment pairs as follows: **i)** Copy 200 original samples and reverse the order of v_1 and v_2. **ii)** Shuffle the 400 Lin samples (the original and reversed samples) with the other ones. **iii)** Evaluate all of the shuffled pairs. Each Lin pair was regarded as correct if either direction was judged correct. In other words, we evaluated the upper bound performance of the LEDIR algorithm.

Table 1 shows the accuracy of the acquired verb entailment pairs for each method. Figure 1 shows the accuracy figures for the N-best entailment pairs for each method, with N being 1,000, 2,000, . . ., or 20,000. We observed the following points from the results. First, $Score$ outperformed all the other methods. Second, $Score$ and Precision, which are directional similarity measures, worked well, while Lin, which is a symmetric one, performed poorly even though the directionality of its output was determined manually.

Method	Acc-1	Acc-2	Acc-3
$Score$	**0.770**	**0.660**	**0.460**
BInc	0.450	0.255	0.125
Precision	0.725	0.545	0.385
Lin	0.590	0.370	0.160
Random	0.050	0.010	0.005

Table 1: Accuracy of verb entailment pairs.

Looking at the evaluated samples, $Score$ successfully acquired pairs in which the entailed verbs generalized entailing verbs that were technical terms. (3) shows examples of $Score$'s outputs.

(3) a. *RSS-haisin-suru* → *todokeru*
RSS-feed-do deliver
"feed the RSS data"

b. *middosippu-maunto-suru* → *tumu*
midship-mounting-do mount
"have (engine) midship-mounted"

The errors made by DIRT (4) and BInc (5) included pairs consisting of technical terms.

(4) *kurakkingu-suru*
software.cracking-do
'crack a (security) system'
→ *koutiku-hosyu-suru*
building-maintenance-do
"build and maintain a system"

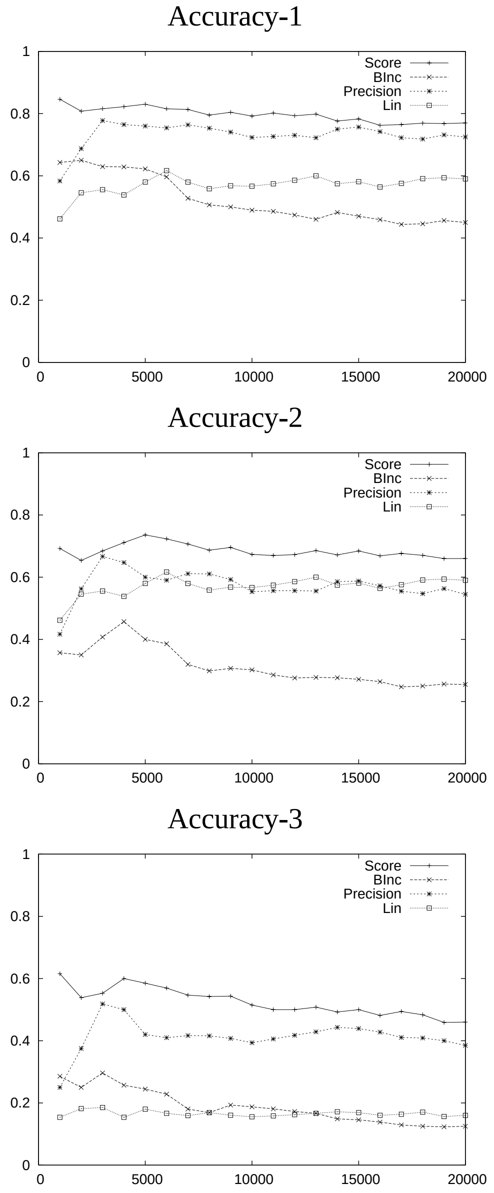

Figure 1: Accuracy of verb entailment pairs.

(5) *suisou-siiku-suru*
tank-raising-do
"raise (fish) in a tank"
→ *siken-houryuu-suru*
test-discharge-do
"stock (with fish) experimentally"

These terms are related in some sense, but they are not entailment pairs.

4.2 Experiment 2: Effectiveness of the Trick

Next, we investigated the effectiveness of the trick described in §3. We evaluated $Score$, $Score_{trick}$, and $Score_{base}$. Table 2 shows the accuracy figures for each method. Figure 2 shows the accuracy figures for the N-best outputs for each method. The

Method	Acc-1	Acc-2	Acc-3
$Score$	**0.770**	**0.660**	**0.460**
$Score_{trick}$	0.725	0.610	0.395
$Score_{base}$	0.590	0.465	0.315

Table 2: Effectiveness of the trick.

results illustrate that introducing the trick significantly improved the performance of $Score_{base}$, and so did multiplying $Score_{trick}$ and $Score_{base}$, which is our proposal $Score$.

(6) shows an example of $Score_{base}$'s errors.

(6)	*gazou-sakusei-suru*	→	*henkou-suru*
	image-making-do		change-do
	"make an image"		"change"

This pair has only two shared nouns ($f \in F_l \cap F_r$), and more than 99.99% of the pair's similarity reflects only one of the two. Clearly, the trick would have prevented the pair from being highly ranked.

4.3 Experiment 3: Pairs of Frequent Verbs

We found that the errors made by Lin and BInc in Experiment 1 were mostly pairs of infrequent verbs such as technical terms. Thus, we conducted the acquisition of entailment pairs targeting more frequent verbs to see how their performance changed. The experimental conditions were the same as in Experiment 1, except that the templates ($\langle p, v\rangle$) used were all $Freq(\langle p, v\rangle) \geq 200$.

Table 3 shows the accuracy figures for each method with the changes in accuracy from those of the original methods in parentheses. The results show that the accuracies of $Score$ and Precision (the two best methods in Experiment 1) degraded, while the other two improved a little. We suspect that the performance difference between these methods would get smaller if we further restricted the target verbs to more frequent ones.

Method	Acc-1	Acc-2	Acc-3
$Score$	0.690 (−0.080)	0.520 (−0.140)	0.335 (−0.125)
BInc	0.455 (+0.005)	0.295 (+0.040)	0.160 (+0.035)
Precision	0.450 (−0.275)	0.355 (−0.190)	0.205 (−0.180)
Lin	0.635 (+0.045)	0.385 (+0.015)	0.205 (+0.045)

Table 3: Accuracy of frequent verb pairs.

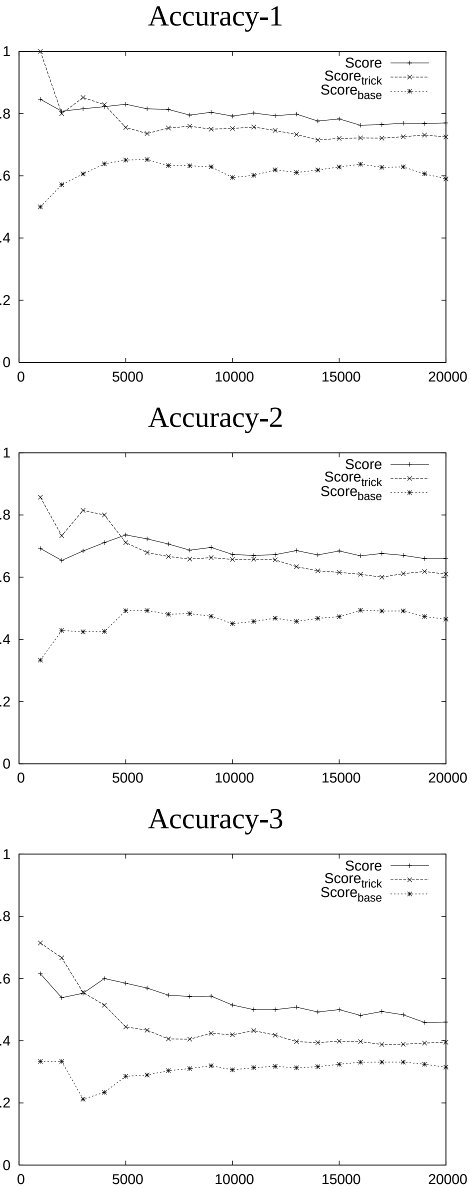

Figure 2: Accuracy of verb entailment pairs acquired by $Score$, $Score_{trick}$, and $Score_{base}$.

However, we believe that dealing with verbs comprehensively, including infrequent ones, is important, since, in the era of information explosion, the impact on applications is determined not only by frequent verbs but also infrequent ones that constitute the long tail of a verb-frequency graph. Thus, this tendency does not matter for our purpose.

4.4 Experiment 4: Template Pairs

This section presents the entailment acquisition accuracy for template pairs to show that our method can also perform the entailment acquisition of unary templates. We presented pairs of unary templates, obtained by the procedure in

§3.4, to the annotators. In doing so, we restricted the correct entailment pairs to those for which entailment always held regardless of what argument filled the two unary slots, and the two slots had to be filled with the same argument, as exemplified in (2). We evaluated $Score$ and Precision.

Table 4 shows the accuracy of the acquired pairs of unary templates. Compared to verb entailment acquisition, the accuracy of both methods dropped by about 10%. This was mainly due to the evaluation restriction exemplified in (2) which was not introduced in the previous experiments; the annotators ignored the argument correspondence between the verb pairs in Experiment 1. Also note that $Score$ outperformed Precision in this experiment, too.

	Method	Acc-1	Acc-2	Acc-3
Scoring Slots	$Score$	**0.655** (−0.115)	**0.510** (−0.150)	**0.300** (−0.160)
	Precision	0.565 (−0.160)	0.430 (−0.115)	0.265 (−0.120)
Nom Slots	$Score$	**0.665** (−0.105)	**0.515** (−0.145)	**0.315** (−0.145)
	Precision	0.490 (−0.235)	0.325 (−0.220)	0.215 (−0.170)

Table 4: Accuracy of entailment pairs of templates whose slots were used for scoring.

(7) and (8) are examples of the Scoring Slots template entailment pairs and (9) is that of the Nom Slots acquired by our method.

(7) *X-wo tatigui-suru* → *X-wo taberu*
X-ACC standing.up.eating-do X-ACC eat
"eat X standing up" "eat X"

(8) *X-de marineedo-suru* → *X-wo ireru*
X-INS marinade-do X-ACC pour
"marinate with X" "pour X"

(9) *X-ga NBA-iri-suru* · · · (was *X-de* (INS))
X-NOM NBA-entering-do
'X joins an NBA team'
→ *X-ga nyuudan-suru* · · · (was *X-de*)
X-NOM enrollment-do
"X joins a team"

4.5 Experiment 5: Verb Pairs form the Top 100,000

Finally, we examined the accuracy of the top 100,000 verb pairs acquired by $Score$ and Precision. As Table 5 shows, $Score$ outperformed Precision. Note also that $Score$ kept a reasonable accuracy for the top 100,000 results (Acc-2: 48%). The accuracy is encouraging enough to consider human annotation for the top 100,000 results to produce a language resource for verb entailment, which we actually plan to do.

Method	Acc-1	Acc-2	Acc-3
$Score$	**0.610**	**0.480**	**0.300**
Precision	0.470	0.295	0.190

Table 5: Accuracy of the top 100,000 verb pairs.

Below are correct verb entailment examples from the top 100,000 results of our method.

(10) The **121**th pair
kaado-kessai-suru → *siharau*
card-payment-do pay
"pay by card" "pay"

(11) The **6,081**th pair
saitei-suru → *sadameru*
adjudicate-do settle
"adjudicate" "settle"

(12) The **15,464**th pair
eraa-syuuryou-suru → *jikkou-suru*
error-termination-do perform-do
"abend" "execute"

(13) The **30,044**th pair
ribuuto-suru → *kidou-suru*
reboot-do start-do
"reboot" "boot"

(14) The **57,653**th pair
rinin-suru → *syuunin-suru*
resignation-do accession-do
"resign" "accede"

(15) The **70,103**th pair
sijou-tounyuu-suru → *happyou-suru*
market-input-do publication-do
"bring to the market" "publicize"

Below are examples of erroneous pairs from our results. (16) is a causal relation but not an entailment. (17) is a contradictory pair.

(16) The **5,475**th pair
juken-suru → *goukaku-suru*
take.an.exam-do acceptance-do
"take an exam" "gain admission"

(17) The **40,504**th pair

ketujou-suru	→	*syutujou-suru*
not.take.part-do		take.part-do
"not take part"		"take part"

5 Conclusion

This paper addressed verb entailment acquisition from the Web, and proposed a novel directional similarity measure $Score$. Through a series of experiments, we showed **i)** that $Score$ outperforms the previously proposed measures, Lin, Precision, and BInc in large scale verb entailment acquisition, **ii)** that our proposed trick implemented in $Score_{trick}$ significantly improves the accuracy of verb entailment acquisition despite its simplicity, **iii)** that $Score$ worked better than the others even when we restricted the target verbs to more frequent ones, **iv)** that our method is also moderately successful at producing template-level entailment pairs, and **v)** that our method maintained reasonable accuracy (in terms of human annotation) for the top 100,000 results. As examples of the acquired verb entailment pairs illustrated, our method can acquire from an ocean of information, namely the Web, a variety of verb entailment pairs ranging from those that are used in daily life to those that are used in very specific fields.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005)*, pages 597–604.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL 2003*, pages 16–23.

Rahul Bhagat, Patrick Pantel, and Eduard Hovy. 2007. Ledir: An unsupervised algorithm for learning directionality of inference rules. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP2007)*, pages 161–170.

Atsushi Fujita and Satoshi Sato. 2008. A probabilistic model for measuring grammaticality and similarity of automatically generated paraphrases of predicate phrases. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING2008)*, pages 225–232.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005)*, pages 107–114.

Ali Ibrahim, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the 2nd International Workshop on Paraphrasing (IWP2003)*, pages 57–64.

Daisuke Kawahara and Sadao Kurohashi. 2006a. Case Frame Compilation from the Web using High-Performance Computing. In *Proceedings of The 5th International Conference on Language Resources and Evaluation (LREC-06)*, pages 1344–1347.

Daisuke Kawahara and Sadao Kurohashi. 2006b. A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL2006)*, pages 176–183.

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

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL1998)*, pages 768–774.

Viktor Pekar. 2006. Acquisition of verb entailment from text. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL2006)*, pages 49–56.

Yusuke Shinyama, Satoshi Sekine, and Kiyoshi Sudo. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of the 2nd international Conference on Human Language Technology Research (HLT2002)*, pages 313–318.

Idan Szpector, Hristo Tanev, Ido Dagan, and Bonaventura Coppola. 2004. Scaling web-based acquisition of entailment relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2004)*, pages 41–48.

Idan Szpektor and Ido Dagan. 2008. Learning entailment rules for unary template. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING2008)*, pages 849–856.

Kentaro Torisawa. 2005. Automatic acquisition of expressions representing preparation and utilization of an object. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP05)*, pages 556–560.

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

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2003)*, pages 81–88.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Maria Teresa Pazienza. 2006. Discovering asymmetric entailment relations between verbs using selectional preferences. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21th International Conference on Computational Linguistics (COLING-ACL2006)*, pages 849–856.