Learning Verb Inference Rules from Linguistically-Motivated Evidence

Hila Weisman[§], **Jonathan Berant**[†], **Idan Szpektor**[‡], **Ido Dagan**[§]
§ Computer Science Department, Bar-Ilan University
† The Blavatnik School of Computer Science, Tel Aviv University
‡ Yahoo! Research Israel
{weismah1,dagan}@cs.biu.ac.il
{jonatha6}@post.tau.ac.il
{idan}@yahoo-inc.com

Abstract

Learning inference relations between verbs is at the heart of many semantic applications. However, most prior work on learning such rules focused on a rather narrow set of information sources: mainly distributional similarity, and to a lesser extent manually constructed verb co-occurrence patterns. In this paper, we claim that it is imperative to utilize information from various textual scopes: verb co-occurrence within a sentence, verb co-occurrence within a document, as well as overall corpus statistics. To this end, we propose a much richer novel set of linguistically motivated cues for detecting entailment between verbs and combine them as features in a supervised classification framework. We empirically demonstrate that our model significantly outperforms previous methods and that information from each textual scope contributes to the verb entailment learning task.

1 Introduction

Inference rules are an important building block of many semantic applications, such as Question Answering (Ravichandran and Hovy, 2002) and Information Extraction (Shinyama and Sekine, 2006). For example, given the sentence "*Churros are coated with sugar*", one can use the rule 'coat → cover' to answer the question "*What are Churros covered with?*". Inference rules specify a directional inference relation between two text fragments, and we follow the *Textual Entailment* modeling of inference (Dagan et al., 2006), which refers to such rules as *entailment rules*. In this work we focus on one of the most important rule types, namely, lexical entailment rules between verbs (*verb entailment*), *e.g.,* 'whisper → talk', 'win → play' and 'buy → own'. The significance of such rules has led to active research in automatic learning of entailment rules between verbs or verb-like structures (Zanzotto et al., 2006; Abe et al., 2008; Schoenmackers et al., 2010).

Most prior efforts to learn verb entailment rules from large corpora employed distributional similarity methods, assuming that verbs are semantically similar if they occur in similar contexts (Lin, 1998; Berant et al., 2012). This led to the automatic acquisition of large scale knowledge bases, but with limited precision. Fewer works, such as *VerbOcean* (Chklovski and Pantel, 2004), focused on identifying verb entailment through verb instantiation of manually constructed patterns. For example, the sentence "*he scared and even startled me*" implies that 'startle → scare'. This led to more precise rule extraction, but with poor coverage since contrary to nouns, in which patterns are common (Hearst, 1992), verbs do not co-occur often within rigid patterns. However, verbs do tend to co-occur in the same document, and also in different clauses of the same sentence.

In this paper, we claim that on top of standard pattern-based and distributional similarity methods, corpus-based learning of verb entailment can greatly benefit from exploiting additional linguistically-motivated cues that are specific to verbs. For instance, when verbs co-occur in different clauses of the same sentence, the syntactic relation between the clauses can be viewed as a proxy for the semantic relation between the verbs. Moreover, we claim that to

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 194–204, Jeju Island, Korea, 12–14 July 2012.

improve performance it is crucial to combine information sources from different textual scopes: verb co-occurrence within a sentence and within a document, distributional similarity over the entire corpus, etc.

Our contribution in this paper is two-fold. First, we suggest a novel set of entailment *indicators* that help to detect the likelihood of verb entailment. Our novel indicators are specific to verbs and are linguistically-motivated. Second, we encode our novel indicators as features within a supervised classification framework and integrate them with other standard features adapted from prior work. This results in a supervised corpus-based learning method that combines verb entailment information at the sentence, document and corpus levels.

We test our model on a manually labeled data set, and show that it outperforms the best performing previous work by 24%. In addition, we examine the effectiveness of indicators that operate at the sentence-level, document-level and corpus-level. This analysis reveals that using a rich and diverse set of indicators that capture sentence-level interactions between verbs substantially improves verb entailment detection.

2 Background

The main approach for learning entailment rules between verbs and verb-like structures has employed the *distributional hypothesis*, which assumes that words with similar meanings appear in similar contexts. For example, we expect the words *'buy'* and *'purchase'* to occur with similar subjects and objects in a large corpus. This observation has led to ample work on developing both symmetric and directional similarity measures that attempt to capture semantic relations between lexical items by comparing their neighborhood context (Lin, 1998; Weeds and Weir, 2003; Geffet and Dagan, 2005; Szpektor and Dagan, 2008; Kotlerman et al., 2010).

A far less explored direction for learning verb entailment involves exploiting verb co-occurrence in a sentence or a document. One prominent work is Chklovsky and Pantel's VerbOcean (2004). In VerbOcean, the authors manually constructed 33 patterns and divided them into five pattern groups, where each group signals one of the following five semantic relations: *similarity*, *strength*, *antonymy*, *enablement* and *happens-before*. For example, the pattern *'Xed and later Yed'* signals the *happens-before* relation between the verbs *'X'* and *'Y'*. Starting with candidate verb pairs based on a distributional similarity measure, the patterns are used to choose a semantic relation per verb pair based on the different patterns this pair instantiates. This method is more precise than distributional similarity approaches, but it is highly susceptible to sparseness issues, since verbs do not typically co-occur within rigid patterns. Utilizing verb co-occurrence at the document level, Chambers and Jurafsky (2008) estimate whether a pair of verbs is narratively related by counting the number of times the verbs share an argument in the same document. In a similar manner, Pekar (2008) detects entailment rules between templates from shared arguments within discourse-related clauses in the same document.

Recently, supervised classification has become standard in performing various semantic tasks. Mirkin et al. (2006) introduced a system for learning entailment rules between nouns (*e.g.,* 'novel → book') that combines distributional similarity and Hearst patterns as features in a supervised classifier. Pennacchiotti and Pantel (2009) augment Mirkin et al's features with web-based features for the task of entity extraction. Hagiwara et al. (2009) perform synonym identification based on both distributional and contextual features. Tremper (2010) extract "loose" sentence-level features in order to identify the *presupposition* relation (*e.g.,* , the verb *'win'* presupposes the verb *'play'*). Last, Berant et al. (2012) utilized various distributional similarity features to identify entailment between lexical-syntactic predicates.

In this paper, we follow the supervised approach for semantic relation detection in order to identify verb entailment. While we utilize and adapt useful features from prior work, we introduce a diverse set of novel features for the task, effectively combining verb co-occurrence information at the sentence, document, and corpus levels.

3 Linguistically-Motivated Indicators

As mentioned in Section 1, verbs behave quite differently from nouns in corpora. In this section, we

introduce linguistically motivated indicators that are specific to verbs and may signal the semantic relation between verb pairs. Then, in Section 4 we describe how these indicators are exactly encoded as features within a supervised classification framework.

Verb co-occurrence When (non-auxiliary) verbs co-occur in a sentence, they are often the main verbs of different clauses. We thus aim to use information about the relation between clauses to learn about the relation between the clauses' main verbs. *Discourse markers* (Hobbs, 1979; Schiffrin, 1988) are lexical terms such as *'because'* and *'however'* that indicate a semantic relation between discourse fragments (i.e., propositions or speech acts). We suggest that these markers can indicate semantic relations between the main verbs of the connected clauses. For example, in the sentence *"He always snores while he sleeps"*, the marker *'while'* indicates a temporal relation between the clauses, indicating that *'snoring'* occurs while *'sleeping'* (and so *'snore* $\rightarrow$ *sleep'*).

Often the relation between clauses is not expressed explicitly with an overt discourse marker, but is still implied by the syntactic structure of the sentence. For example, in dependency parsing the relation can be captured by labeled dependency edges expressing that one clause is an adverbial adjunct of the other, or that two clauses are coordinated. This can indicate the existence (or lack) of entailment between verbs. For instance, in the sentence *"When I walked into the room, he was working out"*, the verb *'walk'* is an adverbial adjunct of the verb *'work out'*. Such co-occurrence structure does not indicate a deep semantic relation, such as entailment, between the two verbs.

Verb classes Verb classes are sets of semantically-related verbs sharing some linguistic properties (Levin, 1993). One of the most general verb classes are stative vs. event verbs (Jackendoff, 1983). Stative verb, such as *'love'* and *'think'*, usually describe a state that lasts some time. On the other hand, event verbs, such as *'run'* and *'kiss'*, describe an action. We hypothesize that verb classes are relevant for determining entailment, for example, that stative verbs are not likely to entail event verbs.

Verb generality Verb-particle constructions are multi-word expressions consisting of a head verb and a particle, *e.g., switch off* (Baldwin and Villavicencio, 2002). We conjecture that the more general a verb is, the more likely it is to appear with many different particles. Detecting verb generality can help us tackle an infamous property of distributional similarity methods, namely, the difficulty in detecting the direction of entailment (Berant et al., 2012). For example, the verb *'cover'* appears with many different particles such as *'up'* and *'for'*, while the verb *'coat'* does not. Thus, assuming we have evidence for an entailment relation between the two verbs, this indicator can help us discern the direction of entailment and determine that 'coat $\rightarrow$ cover'.

Typed Distributional Similarity As discussed in section 2, distributional similarity is the most common source of information for learning semantic relations between verbs. Yet, we suggest that on top of standard distributional similarity measures, which take several verbal arguments into account (such as subject, object, etc.) simultaneously, we should also focus on each type of argument independently. In particular, we apply this approach to compute similarity between verbs based on the set of adverbs that modify them. Our hypothesis is that adverbs may contain relevant information for capturing the direction of entailment. If a verb appears with a small set of adverbs, it is more likely to be a specific verb that already conveys a specific action or state, making an additional adverb redundant. For example, the verb *'whisper'* conveys a specific manner of talking and will probably not appear with the adverb *'loudly'*, while the verb *'talk'* is more likely to appear with such an adverb. Thus, measuring similarity based solely on adverb modifiers could reveal this phenomenon.

4 Supervised Entailment Detection

In the previous section, we discussed linguistic observations regarding novel indicators that may help in detecting entailment relations between verbs. We next describe how to incorporate these indicators as features within a supervised framework for learning lexical entailment rules between verbs. We follow prior work on supervised lexical semantics (Mirkin et al., 2006; Hagiwara et al., 2009; Tremper, 2010)

and address the rule learning task as a classification task. Specifically, given an ordered verb pair (v_1, v_2) as input, we learn a classifier that detects whether the entailment relation '$v_1 \rightarrow v_2$' holds for this pair.

We next detail how our novel indicators, as well as other diverse sources of information found useful in prior work, are encoded as features. Then, we describe the learning model and our feature analysis procedure.

4.1 Entailment features

Most of our features are based on information extracted from the target verb pair co-occurring within varying textual scopes (sentence, document, corpus). Hence, we group the features according to their related scope. Naturally, when the scope is small, *i.e.,* at a sentence level, the semantic relation between the verbs is easier to discern but the information may be sparse. Conversely, when co-occurrence is loose the relation is harder to discern but coverage is increased.

4.1.1 Sentence-level co-occurrence

We next detail features that address co-occurrence of the target verb pair within a sentence. These include our novel linguistically-motivated indicators, as well as features that were adapted from prior work.

Discourse markers As discussed in Section 3, discourse markers may signal relations between the main verbs of adjacent clauses. The literature is abundant with taxonomies that classify markers to various discourse relations (Mann and Thompson, 1988; Hovy and Maier, 1993; Knott and Sanders, 1998). Inspired by Marcu and Echihabi (2002), we employ markers that are mapped to four discourse relations *'Contrast'*, *'Cause'*, *'Condition'* and *'Temporal'*, as specified in Table 1. This definition can be viewed as a relaxed version of VerbOcean's (Chklovski and Pantel, 2004) patterns, although the underlying intuition is different (see Section 3).

For a target verb pair (v_1, v_2) and each discourse relation r, we count the number of times that v_1 is the main verb in the main clause, v_2 is the main verb in the subordinate clause, and the clauses are connected via a marker mapped to r. For example, given the sentence "*You must enroll in the competition before you can participate in it*", the verb pair ('*enroll*','*participate*') appears in the *'Temporal'* relation, indicated by the marker *'before'*, where *'enroll'* is in the main clause. Each count is then normalized by the total number of times (v_1, v_2) appear with any marker. The same procedure is done when v_1 is in the subordinate clause and v_2 in the main clause. We term the features by the relevant discourse relation, *e.g., 'v1-contrast-v2'* refers to v_1 being in the main clause and connected to the subordinate clause via a *contrast* marker.

Dependency relations between clauses As noted in Section 3, the syntactic structure of verb co-occurrence can indicate the existence or lack of entailment. In dependency parsing this may be expressed via the label of the dependency relation connecting the main and subordinate clauses. In our experiments we used the ukWaC corpus[1] (Baroni et al., 2009) which was parsed by the MALT parser (Nivre et al., 2006). Hence, we identified three MALT dependency relations that connect a main clause with its subordinate clause. The first relation is the object complement relation *'obj'*. In this case the subordinate clause is an object complement of the main clause. For example, in "*it surprised me that the lizard could talk*" the verb pair ('*surprise*','*talk*') is connected by the *'obj'* relation. The second relation is the adverbial adjunct relation *'adv'*, in which the subordinate clause is adverbial and describes the time, place, manner, etc. of the main clause, *e.g.,* "*he gave his consent without thinking about the repercussions*". The last relation is the coordination relation *'coord'*, *e.g., "every night my dog Lucky sleeps on the bed **and** my cat Flippers naps in the bathtub*".

Similar to discourse markers, we compute for each verb pair (v_1,v_2) and each dependency label d the proportion of times that v_1 is the main verb of the main clause, v_2 is the main verb of the subordinate clause, and the clauses are connected by dependency relation d, out of all the times they are connected by any dependency relation. We term the features by the dependency label, *e.g., 'v1-adv-v2'* refers to v_1 being in the main clause and connected to the subordinate clause via an *adverbial* adjunct.

[1] http://wacky.sslmit.unibo.it/doku.php?id=corpora

Discourse Rel.	Discourse Markers
Contrast	although , despite , but , whereas , notwithstanding , though
Cause	because , therefore , thus
Condition	if , unless
Temporal	whenever , after , before , until , when , finally , during , afterwards , meanwhile

Table 1: Discourse relations and their mapped markers.

Pattern-based We follow Chklovski and Pantel (2004) and extract occurrences of *VerbOcean* patterns that are instantiated by the target verb pair. As mentioned in Section 2, VerbOcean patterns were originally grouped into five semantic classes. Based on a preliminary study we conducted, we decided to utilize only four *strength*-class patterns as positive indicators for entailment, *e.g.,* "he *scared and even startled* me", and three *antonym*-class patterns as negative indicators for entailment, *e.g.,* "you can *either open or close* the door". We note that these patterns are also commonly used by RTE systems[2].

Since the corpus pattern counts were very sparse, we defined for a target verb pair (v_1, v_2) two binary features: the first denotes whether the verb pair instantiates at least one positive pattern, and the second denotes whether the verb pair instantiates at least one negative pattern. For example, given the aforementioned sentences, the value of the positive feature for the verb pair (*'startle','scare'*) is '1'. Patterns are directional, and so the value of (*'scare','startle'*) is '0'.

Polarity We compute the proportion of times that the two verbs appear in different polarity. For example, in "*he didn't say why he left*", the verb *'say'* appears in negative polarity and the verb *'leave'* in positive polarity. Such change in polarity is usually an indicator of non-entailment between the two verbs.

Tense ordering The temporal relation between verbs may provide information about their semantic relation. For each verb pair co-occurrence, we extract the verbs' tenses and order them as follows: *past* $<$ *present* $<$ *future*. We then add the features *'tense-v1<tense-v2'*, *'tense-v1=tense-v2'*, and *'tense-v1>tense-v2'*, corresponding to the proportion of times the tense of v_1 is smaller, equal to, or bigger than the tense of v_2. This indicates the prevalent temporal relation between the verbs in the corpus and may assist in detecting the direction of entailment. *e.g.,* if tense-v1>tense-v2, the verb pair is less likely to entail.

Co-reference Following Tremper (2010), in every co-occurrence of (v_1,v_2) we extract for each verb the set of arguments at either the subject or object positions, denoted A_1 and A_2 (for v_1 and v_2, respectively). We then compute the proportion of co-occurrences in which v_1 and v_2 share an argument, *i.e.,* $A_1 \cap A_2 \neq \phi$, out of all the co-occurrences in which both A_1 and A_2 are non-empty. The intuition, which is similar to distributional similarity, is that semantically related verbs tend to share arguments.

Syntactic and lexical distance Following Tremper (2010) again, we compute the average distance d in dependency edges between the co-occurring verbs. We compute three features corresponding to three bins indicating if $d < 3$, $3 \leq d \leq 7$, or $d > 7$. Similar features are computed for the distance in words (bins are $0 < d < 5$, $5 \leq d \leq 10$, $d > 10$). This feature provides insight into the syntactic relatedness of the verbs.

Sentence-level pmi Pointwise mutual information (pmi) between v_1 and v_2 is computed, where the co-occurrence scope is a sentence. Higher pmi should hint at semantically related verbs.

4.1.2 Document-level co-occurrence

This group of features addresses co-occurrence of a target verb pair within the same document. These features are less sparse, but tend to capture coarser semantic relations between the target verbs.

Narrative score Chambers and Jurafsky (2008) suggested a method for learning sequences of actions or events (expressed by verbs) in which a sin-

[2] http://aclweb.org/aclwiki/index.php?title=RTE_Knowledge_Resources#Ablation_Tests

gle entity is involved. They proposed a pmi-like *narrative score* (see Eq. (1) in their paper) that estimates whether a pair consisting of a verb and one of its dependency relations (v_1, r_1) is narratively-related to another such pair (v_2, r_2). Their estimation is based on quantifying the likelihood that two verbs will share an argument that instantiates both the dependency position (v_1, r_1) and (v_2, r_2) within documents in which the two verbs co-occur. For example, given the document "*Lindsay was prosecuted for DUI. Lindsay was convicted of DUI.*" the pairs (*'prosecute','subj'*) and (*'convict','subj'*) share the argument *'Lindsay'* and are part of a narrative chain. Such narrative relations may provide cues to the semantic relatedness of the verb pair.

We compute for every target verb pair nine features using their narrative score. In four features, $r_1 = r_2$ and the common dependency is either a subject, an object, a preposition complement (*e.g.,* "we meet at the *station*."), or an adverb (termed *chamb-subj*, *chamb-obj*, and so on). In the next three features, $r_1 \neq r_2$ and r_1, r_2 denote either a subject, object, or preposition complement[3] (termed *chamb-subj-obj* and so on). Last, we add as features the average of the four features where $r_1 = r_2$ (termed *chamb-same*), and the average of the three features where $r_1 \neq r_2$ (termed *chamb-diff*).

Document-level pmi Similar to sentence-level pmi, we compute the pmi between v_1 and v_2, but this time the co-occurrence scope is a document.

4.1.3 Corpus-level statistics

The final group of features ignores sentence or document boundaries and is based on overall corpus statistics.

Distributional similarity Following our hypothesis regarding typed distributional similarity (Section 3), we first compute for each verb and each argument (subject, object, preposition complement and adverb) a separate vector that counts the number of times each word in the corpus instantiates the argument of that verb. In addition, we also compute a vector that is the concatenation of the previous separate vectors, which captures the standard distributional similarity statistics. We then apply three state-of-the-art distributional similarity measures, *Lin* (Lin, 1998), *Weeds precision* (Weeds and Weir, 2003) and *BInc* (Szpektor and Dagan, 2008), to compute for every verb pair a similarity score between each of the five count vectors[4]. We term each feature by the method and argument, *e.g., weeds-prep* and *lin-all* represent the Weeds measure over prepositional complements and the Lin measure over all arguments.

Verb classes Following our discussion in Section 3, we first measure for each target verb v a "stative" feature f by computing the proportion of times it appears in progressive tense, since stative verbs usually do not appear in the progressive tense (*e.g., 'knowing'*). Then, given a verb pair (v_1,v_2) and their corresponding stative features f_1 and f_2, we add two features $f_1 \cdot f_2$ and $\frac{f_1}{f_2}$, which capture the interaction between the verb classes of the two verbs.

Verb generality For each verb, we add as a feature the number of different particles it appears with in the corpus, following the hypothesis that this is a cue to its generality. Then, given a verb pair (v_1,v_2) and their corresponding features f_1 and f_2, we add the feature $\frac{f_1}{f_2}$. We expect that when $\frac{f_1}{f_2}$ is high, v_1 is more general than v_2, which is a negative entailment indicator.

4.2 Learning model and feature analysis

The total number of features in our model as described above is 63. We combine the features in a supervised classification framework with a linear SVM. Since our model contains many novel features, it is important to investigate their utility for detecting verb entailment. To that end, we employ feature ranking methods as suggested by Guyon et al. (2003). In feature ranking methods, features are ranked by some score computed for each feature independently. In this paper we use Pearson correlation between the feature values and the corresponding labels as the ranking criterion.

[3] adverbs never instantiate the subject, object or preposition complement positions.

[4] We employ the common practice of using the pmi between a verb and an argument rather than the argument count as the argument's weight.

5 Evaluation and Analysis

5.1 Experimental Setting

To evaluate our proposed supervised model, we constructed a dataset containing labeled verb pairs. We started by randomly sampling 50 verbs out of the common verbs in the RCV1 corpus[5], which we denote here as *seed verbs*. Next, we extracted the 20 most similar verbs to each seed verb according to the Lin similarity measure (Lin, 1998), which was computed on the RCV1 corpus. Then, for each seed verb v_s and one of its extracted similar verbs v_s^i we generated the two directed pairs (v_s, v_s^i) and (v_s^i, v_s), which represent the candidate rules '$v_s \rightarrow v_s^i$' and '$v_s^i \rightarrow v_s$' respectively. To reduce noise, we filtered out verb pairs where one of the verbs is an auxiliary or a light verb such as *'do'*, *'get'* and *'have'*. This step resulted in 812 verb pairs as our dataset[6], which were manually annotated by the authors as representing a valid entailment rule or not. To annotate these pairs, we generally followed the rule-based approach for entailment rule annotation, where a rule '$v_1 \rightarrow v_2$' is considered as correct if the annotator could think of reasonable contexts under which the rule holds (Dekang and Pantel, 2001; Szpektor et al., 2004). In total 225 verb pairs were labeled as entailing (the rule '$v_1 \rightarrow v_2$' was judged as correct) and 587 verb pairs were labeled as non-entailing (the rule '$v_1 \rightarrow v_2$' was judged as incorrect). The Inter-Annotator Agreement (IAA) for a random sample of 100 pairs was moderate (0.47), as expected from the rule-based approach (Szpektor et al., 2007).

For each verb pair, all 63 features within our model (Section 4) were computed using the ukWaC corpus (Baroni et al., 2009), which contains 2 billion words. For classification, we utilized SVM-perf's (Joachims, 2005) linear SVM implementation with default parameters, and evaluated our model by performing 10-fold cross validation (CV) over the labeled dataset.

[5] http://trec.nist.gov/data/reuters/reuters.html

[6] The data set is available at http://www.cs.biu.ac.il/~nlp/downloads/verb-pair-annotation.html

5.2 Feature selection and analysis

As discussed in Section 4.2, we followed the feature ranking method proposed by Guyon et al. (2003) to investigate the utility of our proposed features. Table 2 depicts the 10 most positively and negatively correlated features with entailment according to the Pearson correlation measure

From Table 2, it is clear that distributional similarity features are amongst the most positively correlated with entailment, which is in line with prior work (Geffet and Dagan, 2005; Kotlerman et al., 2010). Looking more closely, our suggestion for typed distributional similarity proved to be useful, and indeed most of the highly correlated distributional similarity features are typed measures. Standing out are the adverb-typed measures, with two features in the top 10, including the highest, *'Weeds-adverb'*, and *'BInc-adverb'*. We also note that the highly correlated distributional similarity measures are directional, *Weeds* and *BInc*.

The table also indicates that document-level co-occurrence contributes positively to entailment detection. This includes both the Chambers narrative measure, with the typed feature *Chambers-obj*, and document-level PMI, which captures a more loose co-occurrence relationship between verbs. Again, we point at the significant correlation of our novel typed measures with verb entailment, in this case the typed narrative measure.

Last, our feature analysis shows that many of our novel co-occurrence features at the sentence level contribute useful negative information. For example, verbs connected via an adverbial adjunct (*'v2-adverb-v1'*) or an object complement (*'v1-obj-v2'*) are negatively correlated with entailment. In addition, the novel *'verb generality'* feature as well as the tense difference feature (*'tense-v1 > tense-v2'*) are also strong negative indicators. On the other hand, *'v2-coord-v1'* is positively correlated with entailment. This shows that encoding various aspects of verb co-occurrence at the sentence level can lead to better prediction of verb entailment. Finally, we note that PMI at the sentence level is highly correlated with entailment even more than at the document level, since the local textual scope is more indicative, though sparser.

To conclude, our feature analysis shows that fea-

Rank	Top Positive	Top Negative
1	Weeds-adverb	tense-v1 $>$ tense-v2
2	Sentence-level PMI	v2-adverb-v1 co-occurrence
3	Weeds-subj	v2-obj-v1 co-occurrence
4	Weeds-prep	v1-obj-v2 co-occurrence
5	Weeds-all	v1-adverb-v2 co-occurrence
6	Chambers-obj	verb generality $\frac{f_1}{f_2}$
7	v2-coord-v1 co-occurrence	v1-contrast-v2
8	BInc-adverb	tense-v1 $<$ tense-v2
9	Document-level PMI	lexical-distance 0-5
10	Chambers-same	Lin-subj

Table 2: Top 10 positive and negative features according to the Pearson correlation score.

tures at all levels: sentence, document and corpus, contain useful information for entailment detection, both positive and negative, and should be combined together. Moreover, many of our novel features are among the highly correlated features, showing that devising a rich set of verb-specific and linguistically-motivated features provides better discriminative evidence for entailment detection.

5.3 Results and Analysis

We compared our method to the following baselines which were mostly taken from or inspired by prior work:

Random: A simple decision rule: for any pair (v_1, v_2), randomly classify as "yes" with a probability equal to the number of entailing verb pairs out of all verb pairs in the labeled dataset (*i.e.,* $\frac{225}{812} = 0.277$).

VO-KB: A simple unsupervised rule: for any pair (v_1, v_2), classify as "yes" if the pair appears in the *strength* relation (corresponding to entailment) in the VerbOcean knowledge-base, which was computed over Web counts.

VO-ukWaC: A simple unsupervised rule: for any pair (v_1, v_2), classify as "yes" if the value of the positive VerbOcean feature is '1' (Section 4.1, computed over ukWaC).

TDS: Include only the 15 distributional similarity features in our supervised model. This baseline extends Berant et al. (2012), who trained an entailment classifier over several distributional similarity features, and provides an evaluation of the discriminative power of distributional similarity alone, without co-occurrence features.

Method	P%	R%	AUC	F_1
All	**40.2**	**71.0**	**0.65**	**0.51**
TDS+VO	36.8	53.2	0.58	0.41
TDS	34.6	44.8	0.56	0.37
Random	27.9	28.8	0.51	0.28
VO-KB	33.1	14.8	0.53	0.2
VO-ukWaC	23.3	4.7	0.29	0.08

Table 3: Average precision, recall, AUC and F_1 for our method and the baselines.

TDS+VO: Include only the 15 typed distributional similarity features and the two VerbOcean features in our supervised model. This baseline is inspired by Mirkin et al. (2006), who combined distributional similarity features and Hearst patterns (Hearst, 1992) for learning entailment between nouns.

All: Our full-blown model, including all features described in Section 4.1.

For all tested methods, we performed 10-fold cross validation and averaged Precision, Recall, Area under the ROC curve (AUC) and F_1 over the 10 folds. Table 3 presents the results of our full-blown model as well as the baselines.

First, we note that, as expected, the VerbOcean baselines *VO-KB* and *VO-ukWaC* provide low recall,

Method	P%	R%	AUC	F_1
All	40.2	71.0	0.65	0.51
Sent+Corpus-level	39.7	70.4	0.64	0.50
Sent+Doc-level	39.0	70.0	0.63	0.50
Doc+Corpus-level	37.7	64.0	0.62	0.47
Sent-level	35.8	63.8	0.59	0.46
Doc-level	30.0	45.4	0.52	0.35
Corpus-level	35.4	58.1	0.58	0.44

Table 4: Average precision, recall, AUC and F_1 for each subset of the feature groups.

due to the sparseness of rigid pattern instantiation for verbs both in the ukWaC corpus and on the web. Yet, VerbOcean positive and negative patterns do add some discriminative power over only distributional similarity measures, as seen by the improvement of *TDS+VO* over *TDS* in all criteria. But, it is the combination of all types of information sources that yields the best performance. Our complete model, employing the full set of features, outperforms all other models in terms of both precision and recall. Its improvement in terms of F_1 over the second best model (*TDS+VO*), which includes all distributional similarity features as well as pattern-based features, is by 24%. This result shows the benefits of integrating linguistically motivated co-occurrence features with traditional pattern-based and distributional similarity information.

To further investigate the contribution of features at various co-occurrence levels, we trained and tested our model with all possible combinations of feature groups corresponding to a certain co-occurrence scope (sentence, document and corpus). Table 4 presents the results of these tests.

The most notable result of this analysis is that sentence-level features play an important role within our model. Indeed, removing either the document-level features (*Sent+Corpus-level*) or the corpus-level features (*Sent+Doc-level*) results in only a slight decline in performance. Yet, removing the sentence-level features (*Doc+Corpus-level*), ends in a more substantial decline of 8.5% in F_1. In addition, sentence-level features alone (*Sent-level*) provide the best discriminative power for verb entailment, compared to document and corpus levels, which include distributional similarity features. Yet, we note that sentence-level features alone do not capture all the information within our model, and they should be combined with one of the other feature groups to reach performance close to the complete model. This shows again the importance of combining co-occurrence indicators at different levels.

As an additional insight from Table 4, we point out that document-level features are not good entailment indicators by themselves (*Doc-level* in Table 4), and they perform worse than the distributional similarity baseline (*TDS* at Table 3). Still, they do complement each of the other feature groups. In particular, since the *Sent+Doc-level* model performs almost as good as the full model, this subset may be a good substitute to the full model, since its features are easier to extract from large corpora, as they may be extracted in an on-line fashion, processing one document at a time (contrary to corpus-level features).

As a final analysis, we randomly sampled correct entailment rules learned by our model but missed by the typed distributional similarity classifier (TDS). Our overall impression is that employing co-occurrence information helps to better capture entailment relations other than synonymy and troponymy. For example, our model learns that *acquire* $\rightarrow$ *own*, corresponding to the *cause-effect* entailment relation, and that *patent* $\rightarrow$ *invent*, corresponding to the *presupposition* entailment relation.

6 Conclusions and Future Work

We presented a supervised classification model for detecting lexical entailment between verbs. At the heart of our model stand novel linguistically motivated indicators that capture positive and negative entailment information. These indicators encompass co-occurrence relationships between verbs at the sentence, document and corpus level, as well as more fine-grained typed distributional similarity measures. Our model incorporates these novel indicators together with useful features from prior work, combining co-occurrence and distributional similarity information about verb pairs.

Our experiment over a manually labeled dataset showed that our model significantly outperforms several state-of-the-art models both in terms of Pre-

cision and Recall. Further feature analysis indicated that our novel indicators contribute greatly to the performance of the model, and that co-occurrence at multiple levels, combined with distributional similarity features, is necessary to achieve the model's best performance.

In future work we'd like to investigate which indicators may contribute to learning different fine-grained types of entailment, such as presupposition and cause-effect, and attempt to perform a more fine-grained classification to subtypes of entailment.

Acknowledgments

This work was partially supported by the Israel Science Foundation grant 1112/08, the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886, and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287923 (EXCITEMENT).

References

Shuya Abe, Kentaro Inui, and Yuji Matsumoto. 2008. Acquiring event relation knowledge by learning cooccurrence patterns and fertilizing cooccurrence samples with verbal nouns. In *Proceedings of IJCNLP*.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: a case study on verb-particles. In *proceedings of COLING*.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2012. Learning entailment relations by global graph structure optimization. *Computational Linguistics*, 38(1):73–111.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL*.

Timothy Chklovski and Patrick Pantel. 2004. Verb ocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of EMNLP*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Lin Dekang and Patrick Pantel. 2001. Dirt - discovery of inference rules from text. In *In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of ACL*.

Isabelle Guyon and Andre Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiko Toyama. 2009. Supervised synonym acquisition using distributional features and syntactic patterns. In *Journal of Natural Language Processing*.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING*.

Jerry Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3:67–90.

Eduard Hovy and Elisabeth Maier. 1993. *Organizing Discourse Structure Relations using Metafunctions*. Pinter Publishing.

Ray Jackendoff. 1983. *Semantics and Cognition*. The MIT Press.

T. Joachims. 2005. A support vector method for multivariate performance measures. In *Proceedings of ICML*.

Alistair Knott and Ted Sanders. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. In *Journal of Pragmatics*.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University Of Chicago Press.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of ICML*.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of ACL*.

Shachar Mirkin, Ido Dagan, and Maayan Geffet. 2006. Integrating pattern-based and distributional similarity methods for lexical entailment acquisition. In *Proceedings of the COLING/ACL*.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*.

Viktor Pekar. 2008. Discovery of event entailment knowledge from text corpora. *Comput. Speech Lang.*, 22(1):1–16.

Marco Pennacchiotti and Patrick Pantel. 2009. Entity extraction via ensemble semantics. In *Proceedings of EMNLP*.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of ACL*.

Deborah Schiffrin. 1988. *Discourse Markers*. Cambridge University Press.

Stefan Schoenmackers, Jesse Davis, Oren Etzioni, and Daniel S. Weld. 2010. Learning first-order horn clauses from web text. In *Proceedings of EMNLP*.

Yusuke Shinyama and Satoshi Sekine. 2006. Preemptive information extraction using unrestricted relation discovery. In *Proceedings of NAACL-HLT*.

Idan Szpektor and Ido Dagan. 2008. Learning entailment rules for unary templates. In *Proceedings of COLING*.

Idan Szpektor, Hristo Tanev, and Ido Dagan. 2004. Scaling web-based acquisition of entailment relations. In *In Proceedings of EMNLP*.

Idan Szpektor, Eyal Shnarch, and Ido Dagan. 2007. Instance based evaluation of entailment rule acquisition. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Galina Tremper. 2010. Weakly supervised learning of presupposition relations between verbs. In *Proceedings of ACL student workshop*.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of EMNLP*.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Maria Teresa Pazienza. 2006. Discovering asymmetric entailment relations between verbs using selectional preferences. In *Proceedings of the COLING/ACL*.