

# Without a ‘doubt’? Unsupervised discovery of downward-entailing operators

Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Richard Duce

Department of Computer Science

Cornell University

Ithaca, NY 14853-7501

cristian@cs.cornell.edu, llee@cs.cornell.edu, rad47@cornell.edu

## Abstract

An important part of textual inference is making deductions involving *monotonicity*, that is, determining whether a given assertion entails restrictions or relaxations of that assertion. For instance, the statement ‘We *know* the epidemic spread quickly’ does not entail ‘We *know* the epidemic spread quickly via fleas’, but ‘We *doubt* the epidemic spread quickly’ entails ‘We *doubt* the epidemic spread quickly via fleas’. Here, we present the first algorithm for the challenging lexical-semantics problem of learning linguistic constructions that, like ‘doubt’, are *downward entailing* (DE). Our algorithm is unsupervised, resource-lean, and effective, accurately recovering many DE operators that are missing from the hand-constructed lists that textual-inference systems currently use.

Publication venue: NAACL HLT 2009

## 1 Introduction

Making inferences based on natural-language statements is a crucial part of true natural-language understanding, and thus has many important applications. As the field of NLP has matured, there has been a resurgence of interest in creating systems capable of making such inferences, as evidenced by the activity surrounding the ongoing sequence of “Recognizing Textual Entailment” (RTE) competitions (Dagan, Glickman, and Magnini, 2006; Bar-Haim, Dagan, Dolan, Ferro, Giampiccolo, Magnini, and Szpektor, 2006; Giampiccolo, Magnini, Dagan, and Dolan, 2007) and the AQUAINT knowledge-

based evaluation project (Crouch, Saurí, and Fowler, 2005).

The following two examples help illustrate the particular type of inference that is the focus of this paper.

1. ‘We *know* the epidemic spread quickly’
2. ‘We *doubt* the epidemic spread quickly’

A relaxation of ‘*spread quickly*’ is ‘*spread*’; a restriction of it is ‘*spread quickly via fleas*’. From statement 1, we can infer the relaxed version, ‘We *know* the epidemic spread’, whereas the restricted version, ‘We *know* the epidemic spread quickly via fleas’, does not follow. But the reverse holds for statement 2: it entails the restricted version ‘We *doubt* the epidemic spread quickly via fleas’, but not the relaxed version. The reason is that ‘*doubt*’ is a *downward-entailing operator*,<sup>1</sup> in other words, it allows one to, in a sense, “reason from sets to subsets” (van der Wouden, 1997, pg. 90).

Downward-entailing operators are not restricted to assertions about belief or to verbs. For example, the preposition ‘*without*’ is also downward entailing: from ‘*The applicants came without payment or waivers*’ we can infer that all the applicants came without payment. (Contrast this with ‘*with*’, which, like ‘*know*’, is *upward entailing*.) In fact, there are many downward-entailing operators, encompassing many syntactic types; these include explicit negations like ‘*no*’ and ‘*never*’, but also many other

<sup>1</sup>Synonyms for “downward entailing” include *downward-monotonic* and *monotone decreasing*. Related concepts include *anti-additivity*, *veridicality*, and *one-way implicatives*.

terms, such as ‘*refuse (to)*’, ‘*preventing*’, ‘*nothing*’, ‘*rarely*’, and ‘*too [adjective] to*’.

As the prevalence of these operators indicates and as van der Wouden (1997, pg. 92) states, downward entailment “plays an extremely important role in natural language” (van Benthem, 1986; Hoeksema, 1986; Sánchez Valencia, 1991; Dowty, 1994; MacCartney and Manning, 2007). Yet to date, only a few systems attempt to handle the phenomenon in a general way, i.e., to consider more than simple direct negation (Nairn, Condoravdi, and Karttunen, 2006; MacCartney and Manning, 2008; Christodoulopoulos, 2008; Bar-Haim, Berant, Dagan, Greental, Mirkin, Shnarch, and Szpektor, 2008). These systems rely on lists of items annotated with respect to their behavior in “polar” (positive or negative) environments. The lists contain a relatively small number of downward-entailing operators, at least in part because they were constructed mainly by manual inspection of verb lists (although a few non-verbs are sometimes also included). *We therefore propose to automatically learn downward-entailing operators<sup>2</sup> — henceforth DE operators for short — from data; deriving more comprehensive lists of DE operators in this manner promises to substantially enhance the ability of textual-inference systems to handle monotonicity-related phenomena.*

**Summary of our approach** There are a number of significant challenges to applying a learning-based approach. First, to our knowledge there do not exist DE-operator-annotated corpora, and moreover, relevant types of semantic information are “not available in or deducible from any public lexical database” (Nairn et al., 2006). Also, it seems there is no simple test one can apply to all possible candidates; van der Wouden (1997, pg. 110) remarks, “As a rule of thumb, assume that everything that feels negative, and everything that [satisfies a condition

---

<sup>2</sup>We include superlatives (‘*tallest*’), comparatives (‘*taller*’), and conditionals (‘*if*’) in this category because they have non-default (i.e., non-upward entailing) properties — for instance, ‘*he is the tallest father*’ does not entail ‘*he is the tallest man*’. Thus, they also require special treatment when considering entailment relations. In fact, there have been some attempts to unify these various types of non-upward entailing operators (von Fintel, 1999). We use the term *downward entailing (narrowly-defined) (DE(ND))* when we wish to specifically exclude superlatives, comparatives, and conditionals.

described below], is monotone decreasing. This rule of thumb will be shown to be wrong as it stands; but it sort of works, like any rule of thumb.”

Our first insight into how to overcome these challenges is to leverage a finding from the linguistics literature, *Ladusaw’s (1980) hypothesis*, which can be treated as a cue regarding the distribution of DE operators: it asserts that a certain class of lexical constructions known as *negative polarity items (NPIs)* can only appear in the scope of DE operators. Note that this hypothesis suggests that one can develop an *unsupervised* algorithm based simply on checking for co-occurrence with known NPIs.

But there are significant problems with applying this idea in practice, including: (a) there is no agreed-upon list of NPIs; (b) terms can be ambiguous with respect to NPI-hood; and (c) many non-DE operators tend to co-occur with NPIs as well. To cope with these issues, we develop a novel unsupervised *distillation* algorithm that helps filter out the noise introduced by these problems. This algorithm is very effective: it is accurate and derives many DE operators that do not appear on pre-existing lists.

**Contributions** Our project draws a connection between the creation of textual entailment systems and linguistic inquiry regarding DE operators and NPIs, and thus relates to both language-engineering and linguistic concerns.

To our knowledge, this work represents the first attempt to aid in the process of *discovering* DE operators, a task whose importance we have highlighted above. At the very least, our method can be used to provide high-quality raw materials to help human annotators create more extensive DE operator lists. In fact, while previous manual-classification efforts have mainly focused on verbs, we retrieve DE operators across multiple parts of speech. Also, although we discover many items (including verbs) that are not on pre-existing manually-constructed lists, the items we find occur frequently — they are not somehow peculiar or rare.

Our algorithm is surprisingly accurate given that it is quite resource- and knowledge-lean. Specifically, it relies only on Ladusaw’s hypothesis as initial inspiration, a relatively short and arguably noisy list of NPIs, and a large unannotated corpus. It does *not* use other linguistic information — for exam-

ple, we do not use parse information, even though c-command relations have been asserted to play a key role in the licensing of NPIs (van der Wouden, 1997).

## 2 Method

We mentioned in the introduction some significant challenges to developing a machine-learning approach to discovering DE operators. The key insight we apply to surmount these challenges is that in the linguistics literature, it has been hypothesized that there is a strong connection between DE operators and *negative polarity items (NPIs)*, which are terms that tend to occur in “negative environments”. An example NPI is ‘*anymore*’: one can say ‘*We don’t have those anymore*’ but not ‘*\*We have those anymore*’.

Specifically, we propose to take advantage of the seminal hypothesis of Ladusaw (1980, influenced by Fauconnier (1975), *inter alia*):

(Ladusaw) NPIs only appear within the scope of downward-entailing operators.

This hypothesis has been actively discussed, updated, and contested by multiple parties (Linebarger, 1987; von Stechow, 1999; Giannakidou, 2002, *inter alia*). It is not our intent to comment (directly) on its overall validity. Rather, we simply view it as a very useful starting point for developing computational tools to find DE operators—indeed, even detractors of the theory have called it “impressively algorithmic” (Linebarger, 1987, pg. 361).

First, a word about scope. For Ladusaw’s hypothesis, scope should arguably be defined in terms of c-command, immediate scope, and so on (von Stechow, 1999, pg. 100). But for simplicity and to make our approach as resource-lean as possible, we simply assume that potential DE operators occur to the left of NPIs,<sup>3</sup> except that we ignore text to the left of any preceding commas or semi-colons as a way to enforce a degree of locality. For example, in both ‘*By the way, we don’t have plants anymore<sub>NPI</sub> because they died*’ and ‘*we don’t have plants anymore<sub>NPI</sub>*’, we look for DE operators within the sequence of

<sup>3</sup>There are a few exceptions, such as with the NPI “for the life of me” (Hoeksema, 1993).

words ‘*we don’t have plants*’. We refer to such sequences in which we seek DE operators as *NPI contexts*.

Now, Ladusaw’s hypothesis suggests that we can find DE operators by looking for words that tend to occur more often in NPI contexts than they occur overall. We formulate this as follows:

*Assumption:* For any DE operator  $d$ ,  
 $F_{\text{byNPI}}(d) > F(d)$ .

Here,  $F_{\text{byNPI}}(d)$  is the number of occurrences of  $d$  in NPI contexts<sup>4</sup> divided by the number of words in NPI contexts, and  $F(x)$  refers to the number of occurrences of  $x$  relative to the number of words in the corpus.

An additional consideration is that we would like to focus on the discovery of *novel* or non-obvious DE operators. Therefore, for a given candidate DE operator  $c$ , we compute  $\hat{F}_{\text{byNPI}}(c)$ : the value of  $F_{\text{byNPI}}(c)$  that results if we discard all NPI contexts containing a DE operator on a list of 10 well-known instances, namely, ‘*not*’, ‘*n’t*’, ‘*no*’, ‘*none*’, ‘*neither*’, ‘*nor*’, ‘*few*’, ‘*each*’, ‘*every*’, and ‘*without*’. (This list is based on the list of DE operators used by the RTE system presented in MacCartney and Manning (2008).) This yields the following scoring function:

$$S(c) := \frac{\hat{F}_{\text{byNPI}}(c)}{F(c)}. \quad (1)$$

**Distillation** There are certain terms that are not DE operators, but nonetheless co-occur with NPIs as a side-effect of co-occurring with true DE operators themselves. For instance, the proper noun ‘*Milken*’ (referring to Michael Milken, the so-called “junk-bond king”) occurs relatively frequently with the DE operator ‘*denies*’, and ‘*vigorously*’ occurs frequently with DE operators like ‘*deny*’ and ‘*oppose*’. We refer to terms like ‘*milken*’ and ‘*vigorously*’ as “piggybackers”, and address the piggybackers problem by leveraging the following intuition: in general, we do not expect to have two DE operators in the same NPI context.<sup>5</sup> One way to implement this would be

<sup>4</sup>Even if  $d$  occurs multiple times in a single NPI context we only count it once; this way we “dampen the signal” of function words that can potentially occur multiple times in a single sentence.

<sup>5</sup>One reason is that if two DE operators are composed, they ordinarily create a positive context, which would not license

to re-score the candidates in a winner-takes-all fashion: for each NPI context, reward only the candidate with the highest score  $S$ . However, such a method is too aggressive because it would force us to pick a single candidate even when there are several with relatively close scores — and we know our score  $S$  is imperfect. Instead, we propose the following “soft” mechanism. Each sentence distributes a “budget” of total score 1 among the candidates it contains according to the relative scores of those candidates; this works out to yield the following new *distilled* scoring function

$$S_d(c) = \frac{\sum_{\text{NPI contexts } p} \frac{S(c)}{n(p)}}{N(c)}, \quad (2)$$

where  $n(p) = \sum_{c \in p} S(c)$  is an NPI-context normalizing factor and  $N(c)$  is the number of NPI contexts containing the candidate  $c$ . This way, plausible candidates that have high  $S$  scores relative to the other candidates in the sentence receive enhanced  $S_d$  scores. To put it another way: apparently plausible candidates that often appear in sentences with multiple good candidates (i.e., piggybackers) receive a low distilled score, despite a high initial score.

Our general claim is that the higher the distilled score of a candidate, the better its chances of being a DE operator.

**Choice of NPIs** Our proposed method requires access to a set of NPIs. However, there does not appear to be universal agreement on such a set. Lichte and Soehn (2007) mention some doubts regarding approximately 200 (!) of the items on a roughly 350-item list of German NPIs (Kürschner, 1983). For English, the “moderately complete”<sup>6</sup> Lawler (2005) list contains two to three dozen items; however, there is also a list of English NPIs that is several times longer (von Bergen and von Bergen, 1993, written in German), and Hoeksema (1997) asserts that English should have hundreds of NPIs, similarly to French and Dutch.

We choose to focus on the items on these lists that seem most likely to be effective cues for our task. Specifically, we select a subset of the Lawler NPIs, focusing mostly on those that do not have

NPIs (although this is not always the case (Dowty, 1994)).

<sup>6</sup>[www-personal.umich.edu/~jlawler/ae/npi.html](http://www-personal.umich.edu/~jlawler/ae/npi.html)

a relatively frequent non-NPI sense. An example discard is ‘*much*’, whose NPI-hood depends on what it modifies and perhaps on whether there are degree adverbs pre-modifying it (Hoeksema, 1997). There are some ambiguous NPIs that we do retain due to their frequency. For example, ‘*any*’ occurs both in a non-NPI “free choice” variant, as in ‘*any idiot can do that*’, and in an NPI version. Although it is ambiguous with respect to NPI-hood, ‘*any*’ is also a very valuable cue due to its frequency.<sup>7</sup> Here is our NPI list:

any	in weeks/ages/years	budge	yet
at all	drink a drop	red cent	ever
give a damn	last/be/take long	but what	bother to
do a thing	arrive/leave until	give a shit	lift a finger
bat an eye	would care/mind	eat a bite	to speak of

### 3 Experiments

Our main set of evaluations focuses on the precision of our method at discovering new DE operators. We then briefly discuss evaluation of other dimensions.

#### 3.1 Setup

We applied our method to the entirety of the BLLIP (Brown Laboratory for Linguistic Information Processing) 1987–89 WSJ Corpus Release 1, available from the LDC (LDC2000T43). The 1,796,379 sentences in the corpus comprise 53,064 NPI contexts; after discarding the ones containing the 10 well-known DE operators, 30,889 NPI contexts were left. To avoid sparse data problems, we did not consider candidates with very low frequency in the corpus ( $\leq 150$  occurrences) or in the NPI contexts ( $\leq 10$  occurrences).

**Methodology for eliciting judgments** The obvious way to evaluate the precision of our algorithm is to have human annotators judge each output item as to whether it is a DE operator or not. However, there are some methodological issues that arise.

First, if the judges know that every term they are rating comes from our system and that we are hoping that the algorithm extracts DE operators, they may be biased towards calling every item “DE” regardless of whether it actually is. We deal with this problem by introducing *distractors* — items that are not

<sup>7</sup>It is by far the most frequent NPI, appearing in 36,554 of the sentences in the BLLIP corpus (see Section 3).

produced by our algorithm, but are similar enough to not be easily identifiable as “fakes”. Specifically, for each possible part of speech of each of our system’s outputs  $c$  that exists in WordNet, we choose a distractor that is either in a “sibling” synset (a hyponym of  $c$ ’s hypernym) or an antonym. Thus, the distractors are highly related to the candidates. Note that they may in fact also be DE operators.

The judges were made aware of the presence of a substantial number of distractors (about 70 for the set of top 150 outputs). This design choice did seem to help ensure that the judges carefully evaluated each item.

The second issue is that, as mentioned in the introduction, there does not seem to be a uniform test that judges can apply to all items to ascertain their DE-ness; but we do not want the judges to improvise excessively, since that can introduce undesirable randomness into their decisions. We therefore encouraged the judges to try to construct sentences wherein the arguments for candidate DE operators were drawn from a set of phrases and restricted replacements we specified (example: ‘*singing*’ vs ‘*singing loudly*’). However, improvisation was still required in a number of cases; for example, the candidate ‘*act*’, as either a noun or a verb, cannot take ‘*singing*’ as an argument.

The labels that the judges could apply were “DE(ND)” (downward entailing (narrowly-defined)), “superlative”, “comparative”, “conditional”, “hard to tell”, and “not-DE” (= none of the above). We chose this fine-grained sub-division because the second through fourth categories are all known to co-occur with NPIs. There is some debate in the linguistics literature as to whether they can be considered to be downward entailing, narrowly construed, or not (von Fintel, 1999, *inter alia*), but nonetheless, such operators call for special reasoning quite distinct from that required when dealing with upward entailing operators — hence, we consider it a success when our algorithm identifies them.

Since monotonicity phenomena can be rather subtle, the judges engaged in a collaborative process. Judge A (the second author) annotated all items, but worked in batches of around 10 items. At the end of each batch, Judge B (the first author) reviewed Judge A’s decisions, and the two consulted to resolve dis-

agreements as far as possible.

One final remark regarding the annotation: some decisions still seem uncertain, since various factors such as context, Gricean maxims, what should be presupposed<sup>8</sup> and so on come into play. However, we take comfort in a comment by Eugene Charniak (personal communication) to the effect that if a word causes a native speaker to pause, that word is interesting enough to be included. And indeed, it seems reasonable that if a native speaker thinks there might be a sense in which a word can be considered downward entailing, then our system should flag it as a word that an RTE system should at least perhaps pass to a different subsystem for further analysis.

### 3.2 Precision Results

We now examine the 150 items that were most highly ranked by our system, which were subsequently annotated as just described. (For full system output that includes the unannotated items, see <http://www.cs.cornell.edu/~cristian>. We would welcome external annotation help.) As shown in Figure 1a, which depicts precision at  $k$  for various values of  $k$ , our system performs very well. In fact, 100% of the first 60 outputs are DE, broadly construed. It is also interesting to note the increasing presence of instances that the judges found hard to categorize as we move further down the ranking.

Of our 73 distractors, 46% were judged to be members of one of our goal categories. The fact that this percentage is substantially lower than our algorithm’s precision at both 73 and 150 (the largest  $k$  we considered) confirms that our judges were not making random decisions. (We expect the percentage of DE operators among the distractors to be much higher than 0 because they were chosen to be similar to our system’s outputs, and so can be expected to also be DE operators some fraction of the time.)

Table 1 shows the lemmas of just the DE(ND) operators that our algorithm placed in its top 150 outputs.<sup>9</sup> Most of these lemmas are new discoveries, in

<sup>8</sup>For example, ‘*X doubts the epidemic spread quickly*’ might be said to entail ‘*X would doubt the epidemic spreads via fleas, presupposing that X thinks about the flea issue*’.

<sup>9</sup>By listing lemmas, we omit variants of the same word, such as ‘*doubting*’ and ‘*doubted*’, to enhance readability. We omit superlatives, comparatives, and conditionals for brevity.

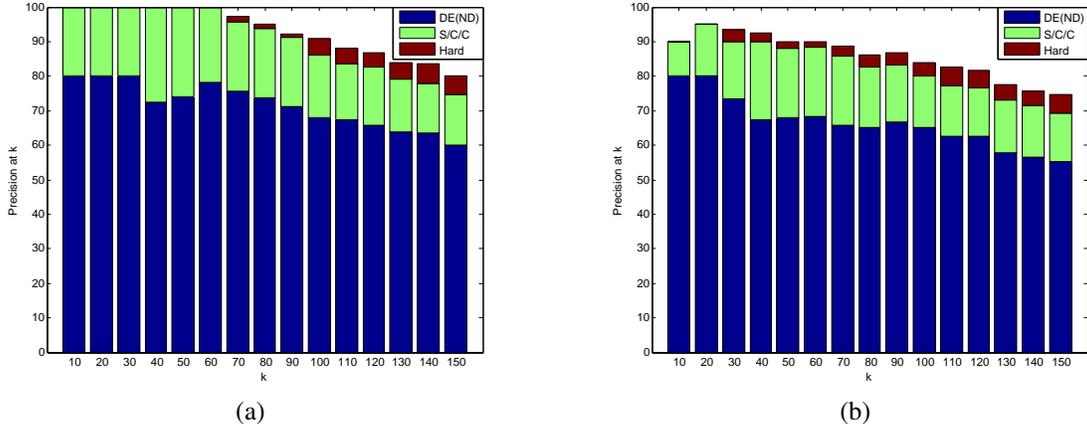


Figure 1: (a) Precision at  $k$  for  $k$  divisible by 10 up to  $k = 150$ . The bar divisions are, from the x-axis up, DE(ND) (blue, the largest); Superlatives/Conditionals/Comparatives (green, 2nd largest); and Hard (red, sometimes non-existent). For example, all of the first 10 outputs were judged to be either downward entailing (narrowly-defined) (8 of 10, or 80%) or in one of the related categories (20%). (b) Precision at  $k$  when the distillation step is omitted.

not-DE			Hard
almost	firmly	one-day	approve
ambitious	fined	signal	cautioned
considers	liable	remove	dismissed
detect	notify	vowed	fend

Table 3: Examples of words judged to be either not in one of our monotonicity categories of interest (not-DE) or hard to evaluate (Hard).

the sense of not appearing in Ladusaw’s (1980) (implicit) enumeration of DE operators. Moreover, the lists of DE(ND) operators that are used by textual-entailment systems are significantly smaller than that depicted in Table 1; for example, MacCartney and Manning (2008) use only about a dozen (personal communication).

Table 3 shows examples of the words in our system’s top 150 outputs that are either clear mistakes or hard to evaluate. Some of these are due to idiosyncrasies of newswire text. For instance, we often see phrases like ‘*biggest one-day drop in ...*’, where ‘*one-day*’ piggybacks on superlatives, and ‘*vowed*’ piggybacks on the DE operator ‘*veto*’, as in the phrase ‘*vowed to veto*’.

**Effect of distillation** In order to evaluate the importance of the distillation process, we study how the results change when distillation is omitted (thus using as score function  $S$  from Equation 1 rather than  $S_d$ ). When comparing the results (summarized in Figure 1b) with those of the complete system

(Figure 1a) we observe that the distillation indeed has the desired effect: the number of highly ranked words that are annotated as not-DE decreases after distillation. This results in an increase of the precision at  $k$  ranging from 5% to 10% (depending on  $k$ ), as can be observed by comparing the height of the composite bars in the two figures.<sup>10</sup>

Importantly, this improvement does indeed seem to stem at least in part from the distillation process handling the piggybacking problem. To give just a few examples: ‘*vigorously*’ is pushed down from rank 48 (undistilled scoring) to rank 126 (distilled scoring), ‘*one-day*’ from 25<sup>th</sup> to 65<sup>th</sup>, ‘*vowed*’ from 45<sup>th</sup> to 75<sup>th</sup>, and ‘*Milken*’ from 121<sup>st</sup> to 350<sup>th</sup>.

### 3.3 Other Results

It is natural to ask whether the (expected) decrease in precision at  $k$  is due to the algorithm assigning relatively low scores to DE operators, so that they do not appear in the top 150, or due to there being no more true DE operators to rank. We cannot directly evaluate our method’s recall because no comprehensive list of DE operators exists. However, to get a rough impression, we can check how our system ranks the items in the largest list we are aware of, namely, the Ladusaw (implicit) list mentioned above. Of the 31 DE operator lemmas on this list (not including the 10 well-known DE operators),

<sup>10</sup>The words annotated “hard” do not affect this increase in precision.

absence of	to defer	hardly (L)	premature to	to rule out	to veto
absent from	to deny (L)	to lack	to prevent	skeptical •	wary of
anxious about •	to deter	innocent of •	to prohibit	to suspend	warned that (L)
to avoid (L)	to discourage	to minimize •	rarely (L)	to thwart	whenever
to bar	to dismiss	never (L)	to refrain from	unable to	withstand
barely	to doubt (L)	nobody	to refuse (L)	unaware of	
to block	to eliminate	nothing	regardless •	unclear on	
cannot (L)	essential for •	to oppose	to reject	unlike	
compensate for •	to exclude	to postpone •	reluctant to (L)	unlikely (L)	
to decline	to fail (L)	to preclude	to resist	unwilling to	

Table 1: The 55 lemmas for the 90 downward entailing (narrowly-defined) operators among our algorithm’s top 150 outputs. (L) marks instances from Ladusaw’s list. • marks some of the more interesting cases. We have added function words (e.g., ‘to’, ‘for’) to indicate parts of speech or subcategorization; our algorithm does not discover multi-word phrases.

Original		Restriction
Dan is <u>unlikely</u> to sing.	$\Rightarrow$ $\Leftarrow$	Dan is <u>unlikely</u> to sing loudly.
Olivia <u>compensates for</u> eating by exercising.	$\Rightarrow$ $\Leftarrow$	Olivia <u>compensates for</u> eating late by exercising.
Ursula <u>refused</u> to sing or dance.	$\Rightarrow$ $\Leftarrow$	Ursula <u>refused</u> to sing.
Bob would <u>postpone</u> singing.	$\Rightarrow$ $\Leftarrow$	Bob would <u>postpone</u> singing loudly.
Talent is <u>essential for</u> singing.	$\Rightarrow$ $\Leftarrow$	Talent is <u>essential for</u> singing a ballad.
She will finish <u>regardless of</u> threats.	$\Rightarrow$ $\Leftarrow$	She will finish <u>regardless of</u> threats to my career.

Table 2: Example demonstrations that the underlined expressions (selected from Table 1) are DE operators: the sentences on the left entail those on the right. We also have provided  $\Leftarrow$  indicators because the reader might find it helpful to reason in the opposite direction and see that these expressions are not upward entailing.

only 7 of those frequent enough to be considered by our algorithm are not in its top 150 outputs, and only 5 are not in the top 300. Remember that we only annotated the top 150 outputs; so, there may be many other DE operators between positions 150 and 300.

Another way of evaluating our method would be to assess the effect of our newly discovered DE operators on downstream RTE system performance. There are two factors to take into account. First, the DE operators we discovered are quite prevalent in naturally occurring text<sup>11</sup>: the 90 DE(ND) operators appearing in our algorithm’s top 150 outputs occur in 111,456 sentences in the BLLIP corpus (i.e., in 6% of its sentences). Second, as previously mentioned, systems do already account for monotonicity to some extent — but they are limited by the fact that their DE operator lexicons are restricted mostly to well-known instances; to take a concrete example with a publicly available RTE system: Nutcracker (Bos and Markert, 2006) correctly infers that ‘We

*did not know the disease spread*’ entails ‘We *did not know the disease spread quickly*’ but it fails to infer that ‘We *doubt the disease spread*’ entails ‘We *doubt the disease spread quickly*’. So, systems can use monotonicity information but currently do not have enough of it; our method can provide them with this information, enabling them to handle a greater fraction of the large number of naturally occurring instances of this phenomenon than ever before.

#### 4 Related work not already discussed

Magnini (2008), in describing modular approaches to textual entailment, hints that NPIs may be used within a negation-detection sub-component.

There is a substantial body of work in the linguistics literature regarding the definition and nature of polarity items (Polarity Items Bibliography). However, very little of this work is computational. There has been passing speculation that one might want to learn polarity-inverting verbs (Christodoulopoulos, 2008, pg. 47). There have also been a few projects on the discovery of NPIs, which is the converse of the problem we consider. Hoeksema (1997)

<sup>11</sup>However, RTE competitions do not happen to currently stress inferences involving monotonicity. The reasons why are beyond the scope of this paper.

discusses some of the difficulties with corpus-based determination of NPIs, including “rampant” polysemy and the problem of “how to determine independently which predicates should count as negative” — a problem which our work addresses. Lichte and Soehn (Lichte, 2005; Lichte and Soehn, 2007) consider finding German NPIs using a method conceptually similar in some respects to our own, although again, their objective is the reverse of ours. Their discovery statistic for single-word NPIs is the ratio of within-licenser-clause occurrences to total occurrences, where, to enhance precision, the list of licensers was filtered down to a set of fairly unambiguous, easily-identified items. They do not consider distillation, which we found to be an important component of our DE-operator-detection algorithm. Their evaluation scheme, unlike ours, did not employ a bias-compensation mechanism. They did employ a collocation-detection technique to extend their list to multi-word NPIs, but our independent experiments with a similar technique (not reported here) did not yield good results.

## 5 Conclusions and future work

To our knowledge, this work represents the first attempt to discover downward entailing operators. We introduced a unsupervised algorithm that is motivated by research in linguistics but employs simple distributional statistics in a novel fashion. Our algorithm is highly accurate and discovers many reasonable DE operators that are missing from pre-existing manually-built lists.

Since the algorithm is resource-lean — requiring no parser or tagger but only a list of NPIs — it can be immediately applied to languages where such lists exist, such as German and Romanian (Trawiński and Soehn, 2008). On the other hand, although the results are already quite good for English, it would be interesting to see what improvements could be gained by using more sophisticated syntactic information.

For languages where NPI lists are not extensive, one could envision applying an iterative co-learning approach: use the newly-derived DE operators to infer new NPIs, and then discover even more new DE operators given the new NPI list. (For English, our initial attempts at bootstrapping from our initial NPI

list on the BLLIP corpus did not lead to substantially improved results.)

In practice, subcategorization is an important feature to capture. In Table 1, we indicate which subcategorizations are DE. An interesting extension of our work would be to try to automatically distinguish particular DE subcategorizations that are lexically apparent, e.g., ‘innocent’ (not DE) vs. ‘innocent of’ (as in ‘innocent of burglary’, DE).

Our project provides a connection (among many) between the creation of textual entailment systems (the domain of language engineers) and the characterization of DE operators (the subject of study and debate among linguists). The prospect that our method might potentially eventually be refined in such a way so as to shed at least a little light on linguistic questions is a very appealing one, although we cannot be certain that any progress will be made on that front.

**Acknowledgments** We thank Roy Bar-Haim, Cleo Condoravdi, and Bill MacCartney for sharing their systems’ lists and information about their work with us; Mats Rooth for helpful conversations; Alex Niculescu-Mizil for technical assistance; and Eugene Charniak for reassuring remarks. We also thank Marisa Ferrara Boston, Claire Cardie, Zhong Chen, Yejin Choi, Effi Georgala, Myle Ott, Stephen Purpura, and Ainur Yessenalina at Cornell University, the UT-Austin NLP group, Roy Bar-Haim, Bill MacCartney, and the anonymous reviewers for their comments on this paper. This paper is based upon work supported in part by DHS grant N0014-07-1-0152, National Science Foundation grant No. BCS-0537606, a Yahoo! Research Alliance gift, a CU Provost’s Award for Distinguished Scholarship, and a CU Institute for the Social Sciences Faculty Fellowship. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of any sponsoring institutions, the U.S. government, or any other entity.

## References

- Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second PASCAL Recognising Textual Entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, 2006.
- Roy Bar-Haim, Jonathan Berant, Ido Dagan, Iddo Green-tal, Shachar Mirkin, Eyal Shnarch, and Idan Szpektor. Efficient semantic deduction and approximate matching over compact parse forests. In *Proceedings of TAC*, 2008.
- Johan Bos and Katja Markert. Recognising textual entailment with robust logical inference. In Quiñonero Candela, Dagan, Magnini, and d’Alché Buc (2006), pages 404–426.

- Christos Christodoulopoulos. Creating a natural logic inference system with combinatory categorial grammar. Master's thesis, University of Edinburgh, 2008.
- Dick Crouch, Roser Saurí, and Abraham Fowler. AQUAINT pilot knowledge-based evaluation: Annotation guidelines. [http://www2.parc.com/istl/groups/nltp/papers/aquaint\\_kb\\_pilot\\_evaluation\\_guide.pdf](http://www2.parc.com/istl/groups/nltp/papers/aquaint_kb_pilot_evaluation_guide.pdf), 2005.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL Recognising Textual Entailment challenge. In Quiñero Candela et al. (2006), pages 177–190.
- David Dowty. The role of negative polarity and concord marking in natural language reasoning. In Mandy Harvey and Lynn Santelmann, editors, *Proceedings of SALT IV*, pages 114–144, Ithaca, New York, 1994. Cornell University.
- Gilles Fauconnier. Polarity and the scale principle. In *Proceedings of the Chicago Linguistic Society (CLS)*, pages 188–199, 1975. Reprinted in Javier Gutierrez-Rexach (ed.), *Semantics: Critical Concepts in Linguistics*, 2003.
- Daniilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL Recognizing Textual Entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, 2007. URL <http://www.aclweb.org/anthology/W/W07/W07-1401>.
- Anastasia Giannakidou. Licensing and sensitivity in polarity items: from downward entailment to nonveridicality. In *Proceedings of the Chicago Linguistic Society (CLS)*, 2002.
- Jack Hoeksema. Monotonicity phenomena in natural language. *Linguistic Analysis*, 16:25–40, 1986.
- Jack Hoeksema. As (of) yet. Appears in *Language and Cognition 3*, the 1992 yearbook of the research group for theoretical and experimental linguistics of the University of Groningen, 1993. <http://www.let.rug.nl/hoeksema/asofyet.pdf>.
- Jack Hoeksema. Corpus study of negative polarity items. *IV-V Jornades de corpus linguistics 1996-1997*, 1997. <http://odur.let.rug.nl/~hoeksema/docs/barcelona.html>.
- Wilfried Kürschner. *Studien zur Negation im Deutschen*. Narr, 1983.
- William A. Ladusaw. *Polarity Sensitivity as Inherent Scope Relations*. Garland Press, New York, 1980. Ph.D. thesis date 1979.
- John Lawler. Negation and NPIs. <http://www.umich.edu/~jlawler/NPIs.pdf>, 2005. Version of 10/29/2005.
- Timm Lichte. Corpus-based acquisition of complex negative polarity items. In *ESSLLI Student Session*, 2005.
- Timm Lichte and Jan-Philipp Soehn. The retrieval and classification of Negative Polarity Items using statistical profiles. In Sam Featherston and Wolfgang Sternefeld, editors, *Roots: Linguistics in Search of its Evidential Base*, pages 249–266. Mouton de Gruyter, 2007.
- Marcia Linebarger. Negative polarity and grammatical representation. *Linguistics and philosophy*, 10:325–387, 1987.
- Bill MacCartney and Christopher D. Manning. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200, 2007.
- Bill MacCartney and Christopher D. Manning. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL <http://www.aclweb.org/anthology/C08-1066>.
- Bernardo Magnini. Slides for a presentation entitled “Semantic Knowledge for Textual Entailment”. Symposium on Semantic Knowledge Discovery, Organization and Use, New York University, November 14 and 15, 2008.
- Rowan Nairn, Cleo Condoravdi, and Lauri Karttunen. Computing relative polarity for textual inference. In *Proceedings of Inference in Computational Semantics (ICoS)*, 2006.
- Polarity Items Bibliography. The polarity items bibliography. <http://www.sfb441.uni-tuebingen.de/a5/pib/XML2HTML/list.html>, 2008. Maintenance guaranteed only through December 2008.
- Joaquin Quiñero Candela, Ido Dagan, Bernardo Magnini, and Florence d’Alché Buc, editors. *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, volume 3944 of *Lecture Notes in Computer Science (LNCS)*, 2006. Springer.
- Víctor Sánchez Valencia. *Studies on natural logic and categorial grammar*. PhD thesis, University of Amsterdam, 1991.
- Beata Trawiński and Jan-Philipp Soehn. A Multilingual Database of Polarity Items. In *Proceedings of LREC 2008, May 28–30, Marrakech, Morocco*, 2008.
- Johan van Benthem. *Essays in Logical Semantics*. Reidel, Dordrecht, 1986.
- Ton van der Wouden. *Negative contexts: Collocation, polarity and multiple negation*. Routledge, 1997.
- Anke von Bergen and Karl von Bergen. *Negative Polarität im Englischen*. Gunter Narr, 1993. List extracted and compiled by Manfred Sailer, 2008, <http://www.sfs.uni-tuebingen.de/~fr/esslli/08/byday/english-npi.pdf>.
- Kai von Fintel. NPI licensing, Strawson entailment, and context dependency. *Journal of Semantics*, 16:97–148, 1999.