Evaluating the Inferential Utility of Lexical-Semantic Resources

Shachar Mirkin, Ido Dagan, Eyal Shnarch
Computer Science Department, Bar-Ilan University
Ramat-Gan 52900, Israel
{mirkins,dagan,shey}@cs.biu.ac.il

Abstract

Lexical-semantic resources are used extensively for applied semantic inference, yet a clear quantitative picture of their current utility and limitations is largely missing. We propose system- and application-independent evaluation and analysis methodologies for resources' performance, and systematically apply them to seven prominent resources. Our findings identify the currently limited recall of available resources, and indicate the potential to improve performance by examining non-standard relation types and by distilling the output of distributional methods. Further, our results stress the need to include auxiliary information regarding the lexical and logical contexts in which a lexical inference is valid, as well as its prior validity likelihood.

1 Introduction

Lexical information plays a major role in semantic inference, as the meaning of one term is often inferred form another. Lexical-semantic resources, which provide the needed knowledge for lexical inference, are commonly utilized by applied inference systems (Giampiccolo et al., 2007) and applications such as Information Retrieval and Question Answering (Shah and Croft, 2004; Pasca and Harabagiu, 2001). Beyond WordNet (Fellbaum, 1998), a wide range of resources has been developed and utilized, including extensions to WordNet (Moldovan and Rus, 2001; Snow et al., 2006) and resources based on automatic distributional similarity methods (Lin, 1998; Pantel and Lin, 2002). Recently, Wikipedia is emerging as a source for extracting semantic relationships (Suchanek et al., 2007; Kazama and Torisawa, 2007).

As of today, only a partial comparative picture is available regarding the actual utility and limitations of available resources for lexical-semantic inference. Works that do provide quantitative information regarding resources utility have focused on few particular resources (Kouylekov and Magnini, 2006; Roth and Sammons, 2007) and evaluated their impact on a specific system. Most often, works which utilized lexical resources do not provide information about their isolated contribution; rather, they only report overall performance for systems in which lexical resources serve as components.

Our paper provides a step towards clarifying this picture. We propose a system- and application-independent evaluation methodology that isolates resources' performance, and systematically apply it to seven prominent lexical-semantic resources. The evaluation and analysis methodology is specified within the Textual Entailment framework, which has become popular in recent years for modeling practical semantic inference in a generic manner (Dagan and Glickman, 2004). To that end, we assume certain definitions that extend the textual entailment paradigm to the lexical level.

The findings of our work provide useful insights and suggested directions for two research communities: developers of applied inference systems and researchers addressing lexical acquisition and resource construction. Beyond the quantitative mapping of resources' performance, our analysis points at issues concerning their effective utilization and major characteristics. Even more importantly, the results highlight current gaps in existing resources and point at directions towards filling them. We show that the coverage of most resources is quite limited, where a substantial part of recall is attributable to semantic relations that are typically not available to inference systems. Notably, distributional acquisition methods

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 558–566,
Athens, Greece, 30 March – 3 April 2009.

are shown to provide many useful relationships which are missing from other resources, but these are embedded amongst many irrelevant ones. Additionally, the results highlight the need to represent and inference over various aspects of contextual information, which affect the applicability of lexical inferences. We suggest that these gaps should be addressed by future research.

2 Sub-sentential Textual Entailment

Textual entailment captures the relation between a *text* t and a textual statement (termed *hypothesis*) h, such that a person reading t would infer that h is most likely correct (Dagan et al., 2005).

The entailment relation has been defined insofar in terms of truth values, assuming that h is a complete sentence (proposition). However, there are major aspects of inference that apply to the sub-sentential level. First, in certain applications, the target hypotheses are often sub-sentential. For example, search queries in IR, which play the hypothesis role from an entailment perspective, typically consist of a single term, like *drug legalization*. Such sub-sentential hypotheses are not regarded naturally in terms of truth values and therefore do not fit well within the scope of the textual entailment definition. Second, many entailment models apply a compositional process, through which they try to infer each sub-part of the hypothesis from some parts of the text (Giampiccolo et al., 2007).

Although inferences over sub-sentential elements are being applied in practice, so far there are no standard definitions for entailment at sub-sentential levels. To that end, and as a prerequisite of our evaluation methodology and our analysis, we first establish two relevant definitions for sub-sentential entailment relations: (a) entailment of a sub-sentential hypothesis by a text, and (b) entailment of one lexical element by another.

2.1 Entailment of Sub-sentential Hypotheses

We first seek a definition that would capture the entailment relationship between a text and a sub-sentential hypothesis. A similar goal was addressed in (Glickman et al., 2006), who defined the notion of *lexical reference* to model the fact that in order to entail a hypothesis, the text has to entail each non-compositional lexical element within it. We suggest that a slight adaptation of their definition is suitable to capture the notion of entailment for any sub-sentential hypotheses, including compositional ones:

Definition 1 *A sub-sentential hypothesis h is entailed by a text t if there is an explicit or implied reference in t to a possible meaning of h.*

For example, the sentence "*crude steel output is likely to fall in 2000*" entails the sub-sentential hypotheses *production*, *steel production* and *steel output decrease*.

Glickman et al., achieving good inter-annotator agreement, empirically found that almost all non-compositional terms in an entailed sentential hypothesis are indeed referenced in the entailing text. This finding suggests that the above definition is consistent with the original definition of textual entailment for sentential hypotheses and can thus model compositional entailment inferences.

We use this definition in our annotation methodology described in Section 3.

2.2 Entailment between Lexical Elements

In the majority of cases, the reference to an "atomic" (non-compositional) lexical element e in h stems from a particular lexical element e' in t, as in the example above where the word *output* implies the meaning of *production*.

To identify this relationship, an entailment system needs a knowledge resource that would specify that the meaning of e' implies the meaning of e, at least in some contexts. We thus suggest the following definition to capture this relationship between e' and e:

Definition 2 *A lexical element e' entails another lexical element e, denoted e'⇒e, if there exist some natural (non-anecdotal) texts containing e' which entail e, such that the reference to the meaning of e can be implied solely from the meaning of e' in the text.*

(Entailment of e by a text follows Definition 1).

We refer to this relation in this paper as *lexical entailment*[1], and call *e' ⇒ e* a *lexical entailment rule*. e' is referred to as the rule's left hand side (*LHS*) and e as its right hand side (*RHS*).

Currently there are no knowledge resources designed specifically for lexical entailment modeling. Hence, the types of relationships they capture do not fully coincide with entailment inference needs. Thus, the definition suggests a specification for the rules that should be provided by

[1] Section 6 discusses other definitions of lexical entailment

a lexical entailment resource, following an operative rationale: a rule *e'* ⇒ *e* should be included in an entailment knowledge resource if it would be needed, as part of a compositional process, to infer the meaning of e from some natural texts. Based on this definition, we perform an analysis of the relationships included in lexical-semantic resources, as described in Section 5.

A rule need not apply in all contexts, as long as it is appropriate for some texts. Two contextual aspects affect rule applicability. First is the "lexical context" specifying the meanings of the text's words. A rules is applicable in a certain context only when the intended sense of its LHS term matches the sense of that term in the text. For example, the application of the rule *lay* ⇒ *produce* is valid only in contexts where the producer is poultry and the products are eggs. This is a well known issue observed, for instance, by Voorhees (1994).

A second contextual factor requiring validation is the "logical context". The logical context determines the monotonicity of the LHS and is induced by logical operators such as negation and (explicit or implicit) quantifiers. For example, the rule *mammal* ⇒ *whale* may not be valid in most cases, but is applicable in universally quantified texts like "*mammals are warm-blooded*". This issue has been rarely addressed in applied inference systems (de Marneffe et al., 2006). The above mentioned rules both comply with Definition 2 and should therefore be included in a lexical entailment resource.

3 Evaluating Entailment Resources

Our evaluation goal is to assess the utility of lexical-semantic resources as sources for entailment rules. An inference system applies a rule by inferring the rule's RHS from texts that match its LHS. Thus, the utility of a resource depends on the performance of its rule applications rather than on the proportion of correct rules it contains. A rule, whether correct or incorrect, has insignificant effect on the resource's utility if it rarely matches texts in real application settings. Additionally, correct rules might produce incorrect applications when applied in inappropriate contexts. Therefore, we use an *instance-based* evaluation methodology, which simulates rule applications by collecting texts that contain rules' LHS and manually assessing the correctness of their applications.

Systems typically handle lexical context either implicitly or explicitly. Implicit context validation occurs when the different terms of a composite hypothesis disambiguate each other. For example, the rule *waterside* ⇒ *bank* is unlikely to be applied when trying to infer the hypothesis *bank loans*, since texts that match *waterside* are unlikely to contain also the meaning of *loan*. Explicit methods, such as word-sense disambiguation or sense matching, validate each rule application according to the broader context in the text. Few systems also address logical context validation by handling quantifiers and negation. As we aim for a system-independent comparison of resources, and explicit approaches are not standardized yet within inference systems, our evaluation uses only implicit context validation.

3.1 Evaluation Methodology

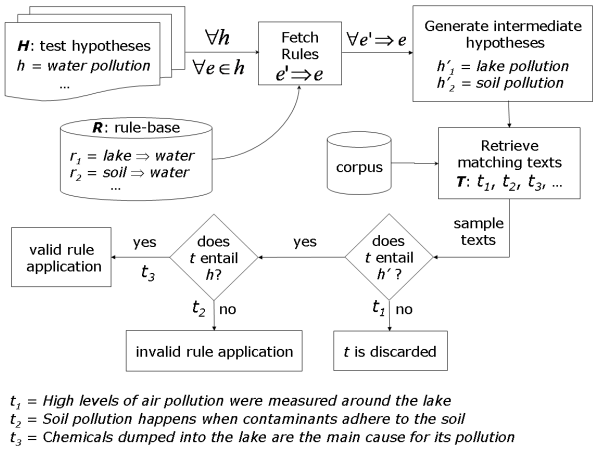

Figure 1: Evaluation methodology flow chart

The input for our evaluation methodology is a lexical-semantic resource R, which contains lexical entailment rules. We evaluate R's utility by testing how useful it is for inferring a sample of test hypotheses H from a corpus. Each hypothesis in H contains more than one lexical element in order to provide implicit context validation for rule applications, e.g. h: *water pollution*.

We next describe the steps of our evaluation methodology, as illustrated in Figure 1. We refer to the examples in the figure when needed:

1) Fetch rules: For each $h \in H$ and each lexical element $e \in h$ (e.g. *water*), we fetch all rules *e'* ⇒ *e* in R that might be applied to entail e (e.g. *lake* ⇒ *water*).

2) Generate intermediate hypotheses h': For each rule r: *e'* ⇒ *e*, we generate an intermediate hypothesis h' by replacing e in h with e' (e.g.

h'_1: *lake pollution*). From a text t entailing h', h can be further entailed by the single application of r. We thus simulate the process by which an entailment system would infer h from t using r.

3) Retrieve matching texts: For each h' we retrieve from a corpus all texts that contain the lemmatized words of h' (not necessarily as a single phrase). These texts may entail h'. We discard texts that also match h since entailing h from them might not require the application of any rule from the evaluated resource. In our example, the retrieved texts contain *lake* and *pollution* but do not contain *water*.

4) Annotation: A sample of the retrieved texts is presented to human annotators. The annotators are asked to answer the following two questions for each text, simulating the typical inference process of an entailment system:

a) Does t entail h'? If t does not entail h' then the text would not provide a useful example for the application of r. For instance, t_1 (in Figure 1) does not entail h'_1 and thus we cannot deduce h from it by applying the rule r. Such texts are discarded from further evaluation.

b) Does t entail h? If t is annotated as entailing h', an entailment system would then infer h from h' by applying r. If h is not entailed from t even though h' is, the rule application is considered invalid. For instance, t_2 does not entail h even though it entails h'_2. Indeed, the application of r_2: **soil* $\Rightarrow$ *water* [2], from which h'_2 was constructed, yields incorrect inference. If the answer is 'yes', as in the case of t_3, the application of r for t is considered valid.

The above process yields a sample of annotated rule applications for each test hypothesis, from which we can measure resources performance, as described in Section 5.

4 Experimental Setting

4.1 Dataset and Annotation

Current available state-of-the-art lexical-semantic resources mainly deal with nouns. Therefore, we used nominal hypotheses for our experiment[3].

We chose TREC 1-8 (excluding 4) as our test corpus and randomly sampled 25 ad-hoc queries of two-word compounds as our hypotheses. We did not use longer hypotheses to ensure that enough texts containing the intermediate hypotheses are found in the corpus. For annotation simplicity, we retrieved single sentences as our texts.

For each rule applied for an hypothesis h, we sampled 10 sentences from the sentences retrieved for that rule. As a baseline, we also sampled 10 sentences for each original hypothesis h in which both words of h are found. In total, 1550 unique sentences were sampled and annotated by two annotators.

To assess the validity of our evaluation methodology, the annotators first judged a sample of 220 sentences. The Kappa scores for inter-annotator agreement were 0.74 and 0.64 for judging h' and h, respectively. These figures correspond to substantial agreement (Landis and Koch, 1997) and are comparable with related semantic annotations (Szpektor et al., 2007; Bhagat et al., 2007).

4.2 Lexical-Semantic Resources

We evaluated the following resources:

WordNet (*WN*^{d}): There is no clear agreement regarding which set of WordNet relations is useful for entailment inference. We therefore took a conservative approach using only synonymy and hyponymy rules, which typically comply with the lexical entailment relation and are commonly used by textual entailment systems, e.g. (Herrera et al., 2005; Bos and Markert, 2006). Given a term e, we created a rule *e'* $\Rightarrow$ *e* for each e' amongst the synonyms or direct hyponyms for all senses of e in WordNet 3.0.

Snow (*Snow*^{30k}): Snow et al. (2006) presented a probabilistic model for taxonomy induction which considers as features paths in parse trees between related taxonomy nodes. They show that the best performing taxonomy was the one adding 30,000 hyponyms to WordNet. We created an entailment rule for each new hyponym added to WordNet by their algorithm[4].

LCC's extended WordNet (*XWN**): In (Moldovan and Rus, 2001) WordNet glosses were transformed into logical form axioms. From this representation we created a rule *e'* $\Rightarrow$ *e* for each e' in the gloss which was tagged as referring to the same entity as e.

CBC: A knowledgebase of labeled clusters generated by the statistical clustering and labeling algorithms in (Pantel and Lin, 2002; Pantel and

[2] The asterisk marks an incorrect rule.

[3] We suggest that the definitions and methodologies can be applied for other parts of speech as well.

[4] Available at http://ai.stanford.edu/~rion/swn

Ravichandran, 2004)[5]. Given a cluster label e, an entailment rule *e'* $\Rightarrow$ *e* is created for each member e' of the cluster.

Lin Dependency Similarity (*Lin-dep*): A distributional word similarity resource based on syntactic-dependency features (Lin, 1998). Given a term e and its list of similar terms, we construct for each e' in the list the rule *e'* $\Rightarrow$ *e*. This resource was previously used in textual entailment engines, e.g. (Roth and Sammons, 2007).

Lin Proximity Similarity (*Lin-prox*): A knowledgebase of terms with their cooccurrence-based distributionally similar terms. Rules are created from this resource as from the previous one[6].

Wikipedia first sentence (*WikiFS*): Kazama and Torisawa (2007) used Wikipedia as an external knowledge to improve Named Entity Recognition. Using the first step of their algorithm, we extracted from the first sentence of each page a noun that appears in a *is-a* pattern referring to the title. For each such pair we constructed a rule *title* $\Rightarrow$ *noun* (e.g. *Michelle Pfeiffer* $\Rightarrow$ *actress*).

The above resources represent various methods for detecting semantic relatedness between words: Manually and semi-automatically constructed (*WN*^{d} and *XWN**, respectively), automatically constructed based on a lexical-syntactic pattern (*WikiFS*), distributional methods (*Lin-dep* and *Lin-prox*) and combinations of pattern-based and distributional methods (*CBC* and *Snow*^{30k}).

5 Results and Analysis

The results and analysis described in this section reveal new aspects concerning the utility of resources for lexical entailment, and experimentally quantify several intuitively-accepted notions regarding these resources and the lexical entailment relation. Overall, our findings highlight where efforts in developing future resources and inference systems should be invested.

5.1 Resources Performance

Each resource was evaluated using two measures - *Precision* and *Recall-share*, macro averaged over all hypotheses. The results achieved for each resource are summarized in Table 1.

[5]Kindly provided to us by Patrick Pantel.

[6]Lin's resources were downloaded from: http://www.cs.ualberta.ca/~lindek/demos.htm

Resource	Precision (%)	Recall-share (%)
Snow^{30k}	56	8
WN^d	55	24
XWN*	51	9
WikiFS	45	7
CBC	33	9
Lin-dep	28	45
Lin-prox	24	36

Table 1: Lexical resources performance

5.1.1 Precision

The *Precision* of a resource R is the percentage of valid rule applications for the resource. It is estimated by the percentage of texts entailing h from those that entail h': $\frac{count_R(entailing\ h=yes)}{count_R(entailing\ h'=yes)}$.

Not surprisingly, resources such as *WN*^{d}, *XWN** or *WikiFS* achieved relatively high precision scores, due to their accurate construction methods. In contrast, Lin's distributional resources are not designed to include lexical entailment relationships. They provide pairs of contextually similar words, of which many have non-entailing relationships, such as co-hyponyms[7] (e.g. **doctor* $\Rightarrow$ *journalist*) or topically-related words, such as **radiotherapy* $\Rightarrow$ *outpatient*. Hence their relatively low precision.

One visible outcome is the large gap between the perceived high accuracy of resources constructed by accurate methods, most notably *WN*^{d}, and their performance in practice. This finding emphasizes the need for instance-based evaluations, which capture the "real" contribution of a resource. To better understand the reasons for this gap we further assessed the three factors that contribute to incorrect applications: incorrect rules, lexical context and logical context (see Section 2.2). This analysis is presented in Table 2.

From Table 2 we see that the gap for accurate resources is mainly caused by applications of correct rules in inappropriate contexts. More interestingly, the information in the table allows us to asses the lexical "context-sensitivity" of resources. When considering only the COR-LEX rules to recalculate resources precision, we find that *Lin-dep* achieves precision of 71% ($\frac{15\%}{15\%+6\%}$), while *WN*^{d} yields only 56% ($\frac{55\%}{55\%+44\%}$). This result indicates that correct *Lin-dep* rules are less sensitive to lexical context, meaning that their prior likelihoods to

[7]a.k.a. *sister terms* or *coordinate terms*

(%)	Invalid Rule Applications				Valid Rule Applications			
	INCOR	COR-LOG	COR-LEX	Total	INCOR	COR-LOG	COR-LEX	Total (P)
WN^d	1	0	44	45	0	0	55	55
WikiFS	13	0	42	55	3	0	42	45
XWN*	19	0	30	49	0	0	51	51
Snow^{30k}	23	0	21	44	0	0	56	56
CBC	51	12	4	67	14	0	19	33
Lin-prox	59	4	13	76	8	3	13	24
Lin-dep	61	5	6	72	9	4	15	28

Table 2: The distribution of invalid and valid rule applications by rule types: incorrect rules (INCOR), correct rules requiring "logical context" validation (COR-LOG), and correct rules requiring "lexical context" matching (COR-LEX). The numbers of each resource's valid applications add up to the resource's precision.

be correct are higher. This is explained by the fact that *Lin-dep*'s rules are calculated across multiple contexts and therefore capture the more frequent usages of words. WordNet, on the other hand, includes many anecdotal rules whose application is rare, and thus is very sensitive to context. Similarly, *WikiFS* turns out to be very context-sensitive. This resource contains many rules for polysemous proper nouns that are scarce in their proper noun sense, e.g. *Captive* $\Rightarrow$ *computer game*. *Snow^{30k}*, when applied with the same calculation, reaches 73%, which explains how it achieved a comparable result to *WN^d*, even though it contains many incorrect rules in comparison to *WN^d*.

5.1.2 Recall

Absolute recall cannot be measured since the total number of texts in the corpus that entail each hypothesis is unknown. Instead, we measure *recall-share*, the contribution of each resource to recall relative to matching only the words of the original hypothesis without any rules. We denote by $yield(h)$ the number of texts that match h directly and are annotated as entailing h. This figure is estimated by the number of sampled texts annotated as entailing h multiplied by the sampling proportion. In the same fashion, for each resource R, we estimate the number of texts entailing h obtained through entailment rules of the resource R, denoted $yield_R(h)$. Recall-share of R for h is the proportion of the yield obtained by the resource's rules relative to the overall yield with and without the rules from R: $\frac{yield_R(h)}{yield(h)+yield_R(h)}$.

From Table 1 we see that along with their relatively low precision, Lin's resources' recall greatly surpasses that of any other resource, including WordNet[8]. The rest of the resources are even inferior to *WN^d* in that respect, indicating their limited utility for inference systems.

As expected, synonyms and hyponyms in WordNet contributed a noticeable portion to recall in all resources. Additional correct rules correspond to hyponyms and synonyms missing from WordNet, many of them proper names and some slang expressions. These rules were mainly provided by *WikiFS* and *Snow^{30k}*, significantly supplementing WordNet, whose *HasInstance* relation is quite partial. However, there are other interesting types of entailment relations contributing to recall. These are discussed in Sections 5.2 and 5.3. Examples for various rule types are found in Table 3.

5.1.3 Valid Applications of Incorrect Rules

We observed that many entailing sentences were retrieved by inherently incorrect rules in the distributional resources. Analysis of these rules reveals they were matched in entailing texts when the LHS has noticeable statistical correlation with another term in the text that does entail the RHS. For example, for the hypothesis *wildlife extinction*, the rule **species* $\Rightarrow$ *extinction* yielded valid applications in contexts about *threatened* or *endangered species*. Has the resource included a rule between the entailing term in the text and the RHS, the entailing text would have been matched without needing the incorrect rule.

These correlations accounted for nearly a third of Lin resources' recall. Nonetheless, in principle, we suggest that such rules, which do not conform with Definition 2, should not be included in a lexical entailment resource, since they also cause invalid rule applications, while the entailing texts they retrieve will hopefully be matched by addi-

[8] A preliminary experiment we conducted showed that recall does not dramatically improve when using the entire hyponymy subtree from WordNet.

Type	Correct Rules	
HYPO	*Shevardnadze* ⇒ *official*	*Snow*^{30k}
ANT	*efficacy* ⇒ *ineffectiveness*	*Lin-dep*
HOLO	*government* ⇒ *official*	*Lin-prox*
HYPER	*arms* ⇒ *gun*	*Lin-prox*
-	*childbirth* ⇒ *motherhood*	*Lin-dep*
-	*mortgage* ⇒ *bank*	*Lin-prox*
-	*Captive* ⇒ *computer*	*WikiFS*
-	*negligence* ⇒ *failure*	*CBC*
-	*beatification* ⇒ *pope*	*XWN**

Type	Incorrect Rules	
CO-HYP	*alcohol* ⇒ *cigarette*	*CBC*
-	*radiotherapy* ⇒ *outpatient*	*Lin-dep*
-	*teen-ager* ⇒ *gun*	*Snow*^{30k}
-	*basic* ⇒ *paper*	*WikiFS*
-	*species* ⇒ *extinction*	*Lin-prox*

Table 3: Examples of lexical resources rules by types. HYPO: hyponymy, HYPER: hypernymy (class entailment of its members), HOLO: holonymy, ANT: antonymy, CO-HYP: co-hyponymy. The non-categorized relations do not correspond to any WordNet relation.

tional correct rules in a more comprehensive resource.

5.2 Non-standard Entailment Relations

An important finding of our analysis is that some less standard entailment relationships have a considerable impact on recall (see Table 3). These rules, which comply with Definition 2 but do not conform to any WordNet relation type, were mainly contributed by Lin's distributional resources and to a smaller degree are also included in *XWN**. In *Lin-dep*, for example, they accounted for approximately a third of the recall.

Among the finer grained relations we identified in this set are topical entailment (e.g. *IBM* as the company entailing the topic *computers*), consequential relationships (*pregnancy* ⇒ *motherhood*) and an entailment of inherent arguments by a predicate, or of essential participants by a scenario description, e.g. *beatification* ⇒ *pope*. A comprehensive typology of these relationships requires further investigation, as well as the identification and development of additional resources from which they can be extracted.

As opposed to hyponymy and synonymy rules, these rules are typically non-substitutable, i.e. the RHS of the rule is unlikely to have the exact same role in the text as the LHS. Many inference systems perform rule-based transformations, substituting the LHS by the RHS. This finding suggests that different methods may be required to utilize such rules for inference.

5.3 Logical Context

WordNet relations other than synonyms and hyponyms, including antonyms, holonyms and hypernyms (see Table 3), contributed a noticeable share of valid rule applications for some resources. Following common practice, these relations are missing by construction from the other resources.

As shown in Table 2 (COR-LOG columns), such relations accounted for a seventh of *Lin-dep*'s valid rule applications, as much as was the contribution of hyponyms and synonyms to this resource's recall. Yet, using these rules resulted with more erroneous applications than correct ones. As discussed in Section 2.2, the rules induced by these relations do conform with our lexical entailment definition. However, a valid application of these rules requires certain logical conditions to occur, which is not the common case. We thus suggest that such rules are included in lexical entailment resources, as long as they are marked properly by their types, allowing inference systems to utilize them only when appropriate mechanisms for handling logical context are in place.

5.4 Rules Priors

In Section 5.1.1 we observed that some resources are highly sensitive to context. Hence, when considering the validity of a rule's application, two factors should be regarded: the actual context in which the rule is to be applied, as well as the rule's prior likelihood to be valid in an arbitrary context. Somewhat indicative, yet mostly indirect, information about rules' priors is contained in some resources. This includes sense ranks in WordNet, SemCor statistics (Miller et al., 1993), and similarity scores and rankings in Lin's resources. Inference systems often incorporated this information, typically as top-*k* or threshold-based filters (Pantel and Lin, 2003; Roth and Sammons, 2007). By empirically assessing the effect of several such filters in our setting, we found that this type of data is indeed informative in the sense that precision increases as the threshold rises. Yet, no specific filters were found to improve results in terms of F1 score (where recall is measured relatively to the yield of the unfiltered resource) due to a significant drop in relative recall. For example, *Lin-*

prox loses more than 40% of its recall when only the top-50 rules for each hypothesis are exploited, and using only the first sense of WN^d costs the resource over 60% of its recall. We thus suggest a better strategy might be to combine the prior information with context matching scores in order to obtain overall likelihood scores for rule applications, as in (Szpektor et al., 2008). Furthermore, resources should include explicit information regarding the prior likelihoods of of their rules.

5.5 Operative Conclusions

Our findings highlight the currently limited recall of available resources for lexical inference. The higher recall of Lin's resources indicates that many more entailment relationships can be acquired, particularly when considering distributional evidence. Yet, available distributional acquisition methods are not geared for lexical entailment. This suggests the need to develop acquisition methods for dedicated and more extensive knowledge resources that would subsume the correct rules found by current distributional methods. Furthermore, substantially better recall may be obtained by acquiring non-standard lexical entailment relationships, as discussed in Section 5.2, for which a comprehensive typology is still needed. At the same time, transformation-based inference systems would need to handle these kinds of rules, which are usually non-substitutable. Our results also quantify and stress earlier findings regarding the severe degradation in precision when rules are applied in inappropriate contexts. This highlights the need for resources to provide explicit information about the suitable lexical and logical contexts in which an entailment rule is applicable. In parallel, methods should be developed to utilize such contextual information within inference systems. Additional auxiliary information needed in lexical resources is the prior likelihood for a given rule to be correct in an arbitrary context.

6 Related Work

Several prior works defined lexical entailment. WordNet's lexical entailment is a relationship between verbs only, defined for propositions (Fellbaum, 1998). Geffet and Dagan (2004) defined *substitutable lexical entailment* as a relation between substitutable terms. We find this definition too restrictive as non-substitutable rules may also be useful for entailment inference. Examples are *breastfeeding* $\Rightarrow$ *baby* and *hospital* $\Rightarrow$ *medical*. Hence, Definition 2 is more broadly applicable for defining the desired contents of lexical entailment resources. We empirically observed that the rules satisfying their definition are a proper subset of the rules covered by our definition. Dagan and Glickman (2004) referred to entailment at the subsentential level by assigning truth values to subpropositional text fragments through their existential meaning. We find this criterion too permissive. For instance, the existence of *country* implies the existence of its *flag*. Yet, the meaning of *flag* is typically not implied by *country*.

Previous works assessing rule application via human annotation include (Pantel et al., 2007; Szpektor et al., 2007), which evaluate acquisition methods for lexical-syntactic rules. They posed an additional question to the annotators asking them to filter out invalid contexts. In our methodology implicit context matching for the full hypothesis was applied instead. Other related instance-based evaluations (Giuliano and Gliozzo, 2007; Connor and Roth, 2007) performed lexical substitutions, but did not handle the non-substitutable cases.

7 Conclusions

This paper provides several methodological and empirical contributions. We presented a novel evaluation methodology for the utility of lexical-semantic resources for semantic inference. To that end we proposed definitions for entailment at subsentential levels, addressing a gap in the textual entailment framework. Our evaluation and analysis provide a first quantitative comparative assessment of the isolated utility of a range of prominent potential resources for entailment rules. We have shown various factors affecting rule applicability and resources performance, while providing operative suggestions to address them in future inference systems and resources.

Acknowledgments

The authors would like to thank Naomi Frankel and Iddo Greental for their excellent annotation work, as well as Roy Bar-Haim and Idan Szpektor for helpful discussion and advice. This work was partially supported by the Negev Consortium of the Israeli Ministry of Industry, Trade and Labor, the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886 and the Israel Science Foundation grant 1095/05.

References

Rahul Bhagat, Patrick Pantel, and Eduard Hovy. 2007. LEDIR: An unsupervised algorithm for learning directionality of inference rules. In *Proceedings of EMNLP-CoNLL*.

J. Bos and K. Markert. 2006. When logical inference helps determining textual entailment (and when it doesn't). In *Proceedings of the Second PASCAL RTE Challenge*.

Michael Connor and Dan Roth. 2007. Context sensitive paraphrasing with a global unsupervised classifier. In *Proceedings of ECML*.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. In *PASCAL Workshop on Learning Methods for Text Understanding and Mining*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In Joaquin Quinonero Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *MLCW*, Lecture Notes in Computer Science.

Marie-Catherine de Marneffe, Bill MacCartney, Trond Grenager, Daniel Cer, Anna Rafferty, and Christopher D. Manning. 2006. Learning to distinguish valid textual entailments. In *Proceedings of the Second PASCAL RTE Challenge*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Maayan Geffet and Ido Dagan. 2004. Feature vector quality and distributional similarity. In *Proceedings of COLING*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of ACL-WTEP Workshop*.

Claudio Giuliano and Alfio Gliozzo. 2007. Instance based lexical entailment for ontology population. In *Proceedings of EMNLP-CoNLL*.

Oren Glickman, Eyal Shnarch, and Ido Dagan. 2006. Lexical reference: a semantic matching subtask. In *Proceedings of EMNLP*.

Jesús Herrera, Anselmo Peñas, and Felisa Verdejo. 2005. Textual entailment recognition based on dependency analysis and wordnet. In *Proceedings of the First PASCAL RTE Challenge*.

Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of EMNLP-CoNLL*.

Milen Kouylekov and Bernardo Magnini. 2006. Building a large-scale repository of textual entailment rules. In *Proceedings of LREC*.

J. R. Landis and G. G. Koch. 1997. The measurements of observer agreement for categorical data. In *Biometrics*, pages 33:159–174.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of HLT*.

Dan Moldovan and Vasile Rus. 2001. Logic form transformation of wordnet and its applicability to question answering. In *Proceedings of ACL*.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of ACM SIGKDD*.

Patrick Pantel and Dekang Lin. 2003. Automatically discovering word senses. In *Proceedings of NAACL*.

Patrick Pantel and Deepak Ravichandran. 2004. Automatically labeling semantic classes. In *Proceedings of HLT-NAACL*.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning inferential selectional preferences. In *Proceedings of HLT*.

Marius Pasca and Sanda M. Harabagiu. 2001. The informative role of wordnet in open-domain question answering. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*.

Dan Roth and Mark Sammons. 2007. Semantic and logical inference model for textual entailment. In *Proceedings of ACL-WTEP Workshop*.

Chirag Shah and Bruce W. Croft. 2004. Evaluating high accuracy retrieval techniques. In *Proceedings of SIGIR*.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING-ACL*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge - unifying wordnet and wikipedia. In *Proceedings of WWW*.

Idan Szpektor, Eyal Shnarch, and Ido Dagan. 2007. Instance-based evaluation of entailment rule acquisition. In *Proceedings of ACL*.

Idan Szpektor, Ido Dagan, Roy Bar-Haim, and Jacob Goldberger. 2008. Contextual preferences. In *Proceedings of ACL*.

Ellen M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of SIGIR*.