Instance-based Evaluation of Entailment Rule Acquisition

Idan Szpektor, Eyal Shnarch, Ido Dagan
Dept. of Computer Science
Bar Ilan University
Ramat Gan, Israel
{szpekti,shey,dagan}@cs.biu.ac.il

Abstract

Obtaining large volumes of inference knowledge, such as entailment rules, has become a major factor in achieving robust semantic processing. While there has been substantial research on learning algorithms for such knowledge, their evaluation methodology has been problematic, hindering further research. We propose a novel evaluation methodology for entailment rules which explicitly addresses their semantic properties and yields satisfactory human agreement levels. The methodology is used to compare two state of the art learning algorithms, exposing critical issues for future progress.

1 Introduction

In many NLP applications, such as Question Answering (QA) and Information Extraction (IE), it is crucial to recognize that a particular target meaning can be inferred from different text variants. For example, a QA system needs to identify that "*Aspirin lowers the risk of heart attacks*" can be inferred from "*Aspirin prevents heart attacks*" in order to answer the question "*What lowers the risk of heart attacks?*". This type of reasoning has been recognized as a core semantic inference task by the generic *textual entailment* framework (Dagan et al., 2006).

A major obstacle for further progress in semantic inference is the lack of broad-scale knowledge-bases for semantic variability patterns (Bar-Haim et al., 2006). One prominent type of inference knowledge representation is inference rules such as paraphrases and *entailment rules*. We define an entailment rule to be a directional relation between two *templates*, text patterns with variables, e.g. 'X prevent $Y \rightarrow X$ lower the risk of Y'. The left-hand-side template is assumed to entail the right-hand-side template in certain contexts, under the same variable instantiation. Paraphrases can be viewed as bidirectional entailment rules. Such rules capture basic inferences and are used as building blocks for more complex entailment inference. For example, given the above rule, the answer "*Aspirin*" can be identified in the example above.

The need for large-scale inference knowledge-bases triggered extensive research on automatic acquisition of paraphrase and entailment rules. Yet the current precision of acquisition algorithms is typically still mediocre, as illustrated in Table 1 for DIRT (Lin and Pantel, 2001) and TEASE (Szpektor et al., 2004), two prominent acquisition algorithms whose outputs are publicly available. The current performance level only stresses the obvious need for satisfactory evaluation methodologies that would drive future research.

The prominent approach in the literature for evaluating rules, termed here the *rule-based* approach, is to present the rules to human judges asking whether each rule is correct or not. However, it is difficult to explicitly define when a learned rule should be considered correct under this methodology, and this was mainly left undefined in previous works. As the criterion for evaluating a rule is not well defined, using this approach often caused low agreement between human judges. Indeed, the standards for evaluation in this field are lower than other fields: many papers

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 456–463,
Prague, Czech Republic, June 2007.

don't report on human agreement at all and those that do report rather low agreement levels. Yet it is crucial to reliably assess rule correctness in order to measure and compare the performance of different algorithms in a replicable manner. Lacking a good evaluation methodology has become a barrier for further advances in the field.

In order to provide a well-defined evaluation methodology we first explicitly specify when entailment rules should be considered correct, following the spirit of their usage in applications. We then propose a new *instance-based* evaluation approach. Under this scheme, judges are not presented only with the rule but rather with a sample of sentences that match its left hand side. The judges then assess whether the rule holds under each specific example. A rule is considered correct only if the percentage of examples assessed as correct is sufficiently high.

We have experimented with a sample of input verbs for both DIRT and TEASE. Our results show significant improvement in human agreement over the rule-based approach. It is also the first comparison between such two state-of-the-art algorithms, which showed that they are comparable in precision but largely complementary in their coverage.

Additionally, the evaluation showed that both algorithms learn mostly one-directional rules rather than (symmetric) paraphrases. While most NLP applications need directional inference, previous acquisition works typically expected that the learned rules would be paraphrases. Under such an expectation, unidirectional rules were assessed as incorrect, underestimating the true potential of these algorithms. In addition, we observed that many learned rules are context sensitive, stressing the need to learn contextual constraints for rule applications.

2 Background: Entailment Rules and their Evaluation

2.1 Entailment Rules

An entailment rule '$L \rightarrow R$' is a directional relation between two templates, L and R. For example, 'X acquire $Y \rightarrow X$ own Y' or 'X beat $Y \rightarrow X$ play against Y'. Templates correspond to text fragments with variables, and are typically either linear phrases or parse sub-trees.

The goal of entailment rules is to help applications infer one text variant from another. A rule can be applied to a given text only when L can be inferred from it, with appropriate variable instantiation. Then, using the rule, the application deduces that R can also be inferred from the text under the same variable instantiation. For example, the rule 'X lose to $Y \rightarrow Y$ beat X' can be used to infer "*Liverpool beat Chelsea*" from "*Chelsea lost to Liverpool in the semifinals*".

Input	**Correct**	**Incorrect**
X change Y (DIRT)	($\leftrightarrow$) X modify Y	X adopt Y
	($\leftarrow$) X amend Y	X create Y
	($\leftarrow$) X revise Y	X stick to Y
X change Y (TEASE)	($\leftrightarrow$) X alter Y	X maintain Y
	($\rightarrow$) X affect Y	X follow Y
	($\leftarrow$) X extend Y	X use Y

Table 1: Examples of templates suggested by DIRT and TEASE as having an entailment relation, in some direction, with the input template 'X change Y'. The entailment direction arrows were judged manually and added for readability.

Entailment rules should typically be applied only in specific contexts, which we term *relevant contexts*. For example, the rule 'X acquire $Y \rightarrow X$ buy Y' can be used in the context of 'buying' events. However, it shouldn't be applied for "*Students acquired a new language*". In the same manner, the rule 'X acquire $Y \rightarrow X$ learn Y' should be applied only when Y corresponds to some sort of knowledge, as in the latter example.

Some existing entailment acquisition algorithms can add contextual constraints to the learned rules (Sekine, 2005), but most don't. However, NLP applications usually implicitly incorporate some contextual constraints when applying a rule. For example, when answering the question "*Which companies did IBM buy?*" a QA system would apply the rule 'X acquire $Y \rightarrow X$ buy Y' correctly, since the phrase "IBM acquire X" is likely to be found mostly in relevant economic contexts. We thus expect that an evaluation methodology should consider context relevance for entailment rules. For example, we would like both 'X acquire $Y \rightarrow X$ buy Y' and 'X acquire $Y \rightarrow X$ learn Y' to be assessed as correct (the second rule should not be deemed incorrect just because it is not applicable in frequent economic contexts).

Finally, we highlight that the common notion of "paraphrase rules" can be viewed as a special case of entailment rules: a paraphrase '$L \leftrightarrow R$' holds if both templates entail each other. Following the textual entailment formulation, we observe that many applied inference settings require only directional entailment, and a requirement for symmetric paraphrase is usually unnecessary. For example, in order to answer the question "*Who owns Overture?*" it suffices to use a directional entailment rule whose right hand side is 'X own Y', such as 'X acquire $Y \rightarrow X$ own Y', which is clearly not a paraphrase.

2.2 Evaluation of Acquisition Algorithms

Many methods for automatic acquisition of rules have been suggested in recent years, ranging from distributional similarity to finding shared contexts (Lin and Pantel, 2001; Ravichandran and Hovy, 2002; Shinyama et al., 2002; Barzilay and Lee, 2003; Szpektor et al., 2004; Sekine, 2005). However, there is still no common accepted framework for their evaluation. Furthermore, all these methods learn rules as pairs of templates $\{L, R\}$ in a symmetric manner, without addressing rule directionality. Accordingly, previous works (except (Szpektor et al., 2004)) evaluated the learned rules under the paraphrase criterion, which underestimates the practical utility of the learned rules (see Section 2.1).

One approach which was used for evaluating automatically acquired rules is to measure their contribution to the performance of specific systems, such as QA (Ravichandran and Hovy, 2002) or IE (Sudo et al., 2003; Romano et al., 2006). While measuring the impact of learned rules on applications is highly important, it cannot serve as the primary approach for evaluating acquisition algorithms for several reasons. First, developers of acquisition algorithms often do not have access to the different applications that will later use the learned rules as generic modules. Second, the learned rules may affect individual systems differently, thus making observations that are based on different systems incomparable. Third, within a complex system it is difficult to assess the exact quality of entailment rules independently of effects of other system components.

Thus, as in many other NLP learning settings, a direct evaluation is needed. Indeed, the prominent approach for evaluating the quality of rule acquisition algorithms is by human judgment of the learned rules (Lin and Pantel, 2001; Shinyama et al., 2002; Barzilay and Lee, 2003; Pang et al., 2003; Szpektor et al., 2004; Sekine, 2005). In this evaluation scheme, termed here the *rule-based* approach, a sample of the learned rules is presented to the judges who evaluate whether each rule is correct or not. The criterion for correctness is not explicitly described in most previous works. By the common view of context relevance for rules (see Section 2.1), a rule was considered correct if the judge could think of reasonable contexts under which it holds.

We have replicated the rule-based methodology but did not manage to reach a 0.6 Kappa agreement level between pairs of judges. This approach turns out to be problematic because the rule correctness criterion is not sufficiently well defined and is hard to apply. While some rules might obviously be judged as correct or incorrect (see Table 1), judgment is often more difficult due to context relevance. One judge might come up with a certain context that, to her opinion, justifies the rule, while another judge might not imagine that context or think that it doesn't sufficiently support rule correctness. For example, in our experiments one of the judges did not identify the valid "religious holidays" context for the correct rule 'X observe $Y \rightarrow X$ celebrate Y'. Indeed, only few earlier works reported inter-judge agreement level, and those that did reported rather low Kappa values, such as 0.54 (Barzilay and Lee, 2003) and 0.55 - 0.63 (Szpektor et al., 2004).

To conclude, the prominent rule-based methodology for entailment rule evaluation is not sufficiently well defined. It results in low inter-judge agreement which prevents reliable and consistent assessments of different algorithms.

3 Instance-based Evaluation Methodology

As discussed in Section 2.1, an evaluation methodology for entailment rules should reflect the expected validity of their application within NLP systems. Following that line, an entailment rule '$L \rightarrow R$' should be regarded as *correct* if in all (or at least most) relevant contexts in which the instantiated template L is inferred from the given text, the instan-

	Rule	Sentence	Judgment
1	X seek $Y \rightarrow X$ disclose Y	If he is arrested, **he** can immediately seek **bail**.	Left not entailed
2	X clarify $Y \rightarrow X$ prepare Y	**He** didn't clarify **his position on the subject**.	Left not entailed
3	X hit $Y \rightarrow X$ approach Y	**Other earthquakes** have hit **Lebanon** since '82.	Irrelevant context
4	X lose $Y \rightarrow X$ surrender Y	**Bread** has recently lost **its subsidy**.	Irrelevant context
5	X regulate $Y \rightarrow X$ reform Y	**The SRA** regulates **the sale of sugar**.	No entailment
6	X resign $Y \rightarrow X$ share Y	**Lopez** resigned **his post at VW** last week.	No entailment
7	X set $Y \rightarrow X$ allow Y	**The committee** set **the following refunds**.	Entailment holds
8	X stress $Y \rightarrow X$ state Y	**Ben Yahia** also stressed **the need for action**.	Entailment holds

Table 2: Rule evaluation examples and their judgment.

tiated template R is also inferred from the text. This reasoning corresponds to the common definition of entailment in semantics, which specifies that a text L entails another text R if R is true in every circumstance (possible world) in which L is true (Chierchia and McConnell-Ginet, 2000).

It follows that in order to assess if a rule is correct we should judge whether R is typically entailed from those sentences that entail L (within relevant contexts for the rule). We thus present a new evaluation scheme for entailment rules, termed the *instance-based* approach. At the heart of this approach, human judges are presented not only with a rule but rather with a sample of examples of the rule's usage. Instead of thinking up valid contexts for the rule the judges need to assess the rule's validity under the given context in each example. The essence of our proposal is a (apparently non-trivial) protocol of a sequence of questions, which determines rule validity in a given sentence.

We shall next describe how we collect a sample of examples for evaluation and the evaluation process.

3.1 Sampling Examples

Given a rule '$L \rightarrow R$', our goal is to generate evaluation examples by finding a sample of sentences from which L is entailed. We do that by automatically retrieving, from a given corpus, sentences that match L and are thus likely to entail it, as explained below.

For each example sentence, we automatically extract the arguments that instantiate L and generate two phrases, termed *left phrase* and *right phrase*, which are constructed by instantiating the left template L and the right template R with the extracted arguments. For example, the left and right phrases generated for example 1 in Table 2 are "*he seek bail*" and "*he disclose bail*", respectively.

Finding sentences that match L can be performed at different levels. In this paper we match lexical-syntactic templates by finding a sub-tree of the sentence parse that is identical to the template structure. Of course, this matching method is not perfect and will sometimes retrieve sentences that do not entail the left phrase for various reasons, such as incorrect sentence analysis or semantic aspects like negation, modality and conditionals. See examples 1-2 in Table 2 for sentences that syntactically match L but do not entail the instantiated left phrase. Since we should assess R's entailment only from sentences that entail L, such sentences should be ignored by the evaluation process.

3.2 Judgment Questions

For each example generated for a rule, the judges are presented with the given sentence and the left and right phrases. They primarily answer two questions that assess whether entailment holds in this example, following the semantics of entailment rule application as discussed above:

$\mathbf{Q_{le}}$**:** Is the left phrase entailed from the sentence? A positive/negative answer corresponds to a '**Left entailed/not entailed**' judgment.

$\mathbf{Q_{re}}$**:** Is the right phrase entailed from the sentence? A positive/negative answer corresponds to an '**Entailment holds/No entailment**' judgment.

The first question identifies sentences that do not entail the left phrase, and thus should be ignored when evaluating the rule's correctness. While inappropriate matches of the rule left-hand-side may happen and harm an overall system precision, such errors should be accounted for a system's rule matching module rather than for the rules' precision. The second question assesses whether the rule application is valid or not for the current example. See examples 5-8 in Table 2 for cases where entailment does or doesn't hold.

Thus, the judges focus only on the given sentence in each example, so the task is actually to evaluate whether *textual entailment* holds between the sentence (*text*) and each of the left and right phrases (*hypotheses*). Following past experience in textual entailment evaluation (Dagan et al., 2006) we expect a reasonable agreement level between judges.

As discussed in Section 2.1, we may want to ignore examples whose context is irrelevant for the rule. To optionally capture this distinction, the judges are asked another question:

$\mathbf{Q_{rc}}$**:** Is the right phrase a likely phrase in English? A positive/negative answer corresponds to a '**Relevant/Irrelevant context**' evaluation.

If the right phrase is not likely in English then the given context is probably irrelevant for the rule, because it seems inherently incorrect to infer an implausible phrase. Examples 3-4 in Table 2 demonstrate cases of irrelevant contexts, which we may choose to ignore when assessing rule correctness.

3.3 Evaluation Process

For each example, the judges are presented with the three questions above in the following order: (1) $\mathbf{Q_{le}}$ (2) $\mathbf{Q_{rc}}$ (3) $\mathbf{Q_{re}}$. If the answer to a certain question is negative then we do not need to present the next questions to the judge: if the left phrase is not entailed then we ignore the sentence altogether; and if the context is irrelevant then the right phrase cannot be entailed from the sentence and so the answer to $\mathbf{Q_{re}}$ is already known as negative.

The above entailment judgments assume that we can actually ask whether the left or right phrases are correct given the sentence, that is, we assume that a truth value can be assigned to both phrases. This is the case when the left and right templates correspond, as expected, to semantic relations. Yet sometimes learned templates are (erroneously) not relational, e.g. 'X, Y, IBM' (representing a list). We therefore let the judges initially mark rules that include such templates as non-relational, in which case their examples are not evaluated at all.

3.4 Rule Precision

We compute the precision of a rule by the percentage of examples for which entailment holds out of all "relevant" examples. We can calculate the precision in two ways, as defined below, depending on whether we ignore irrelevant contexts or not (obtaining lower precision if we don't). When systems answer an information need, such as a query or question, irrelevant contexts are sometimes not encountered thanks to additional context which is present in the given input (see Section 2.1). Thus, the following two measures can be viewed as upper and lower bounds for the expected precision of the rule applications in actual systems:

upper bound precision: $\frac{\#\textbf{Entailment holds}}{\#\textbf{Relevant context}}$

lower bound precision: $\frac{\#\textbf{Entailment holds}}{\#\textbf{Left entailed}}$

where # denotes the number of examples with the corresponding judgment.

Finally, we consider a rule to be correct only if its precision is at least 80%, which seems sensible for typical applied settings. This yields two alternative sets of correct rules, corresponding to the upper bound and lower bound precision measures. Even though judges may disagree on specific examples for a rule, their judgments may still agree overall on the rule's correctness. We therefore expect the agreement level on rule correctness to be higher than the agreement on individual examples.

4 Experimental Settings

We applied the instance-based methodology to evaluate two state-of-the-art unsupervised acquisition algorithms, DIRT (Lin and Pantel, 2001) and TEASE (Szpektor et al., 2004), whose output is publicly available. DIRT identifies semantically related templates in a local corpus using distributional similarity over the templates' variable instantiations. TEASE acquires entailment relations from the Web for a given input template I by identifying characteristic variable instantiations shared by I and other templates.

For the experiment we used the published DIRT and TEASE knowledge-bases[1]. For every given input template I, each knowledge-base provides a list of learned output templates $\{O_j\}_1^{n_I}$, where n_I is the number of output templates learned for I. Each output template is suggested as holding an entailment relation with the input template I, but the algorithms do not specify the entailment direction(s). Thus, each pair $\{I, O_j\}$ induces two candidate directional entailment rules: '$I \rightarrow O_j$' and '$O_j \rightarrow I$'.

4.1 Test Set Construction

The test set construction consists of three sampling steps: selecting a set of input templates for the two algorithms, selecting a sample of output rules to be evaluated, and selecting a sample of sentences to be judged for each rule.

First, we randomly selected 30 transitive verbs out of the 1000 most frequent verbs in the Reuters RCV1 corpus[2]. For each verb we manually constructed a lexical-syntactic input template by adding subject and object variables. For example, for the verb 'seek' we constructed the template '$X \xleftarrow{subj}$ seek $\xrightarrow{obj} Y$'.

Next, for each input template I we considered the learned templates $\{O_j\}_1^{n_I}$ from each knowledge-base. Since DIRT has a long tail of templates with a low score and very low precision, DIRT templates whose score is below a threshold of 0.1 were filtered out[3]. We then sampled 10% of the templates in each output list, limiting the sample size to be between 5-20 templates for each list (thus balancing between sufficient evaluation data and judgment load). For each sampled template O we evaluated both directional rules, '$I \rightarrow O$' and '$O \rightarrow I$'. In total, we sampled 380 templates, inducing 760 directional rules out of which 754 rules were unique.

Last, we randomly extracted a sample of example sentences for each rule '$L \rightarrow R$' by utilizing a search engine over the first CD of Reuters RCV1. First, we retrieved all sentences containing all lexical terms within L. The retrieved sentences were parsed using the Minipar dependency parser (Lin, 1998), keeping only sentences that syntactically match L (as explained in Section 3.1). A sample of 15 matching sentences was randomly selected, or all matching sentences if less than 15 were found. Finally, an example for judgment was generated from each sampled sentence and its left and right phrases (see Section 3.1). We did not find sentences for 108 rules, and thus we ended up with 646 unique rules that could be evaluated (with 8945 examples to be judged).

4.2 Evaluating the Test-Set

Two human judges evaluated the examples. We randomly split the examples between the judges. 100 rules (1287 examples) were cross annotated for agreement measurement. The judges followed the procedure in Section 3.3 and the correctness of each rule was assessed based on both its upper and lower bound precision values (Section 3.4).

5 Methodology Evaluation Results

We assessed the instance-based methodology by measuring the agreement level between judges. The judges agreed on 75% of the 1287 shared examples, corresponding to a reasonable Kappa value of 0.64. A similar kappa value of 0.65 was obtained for the examples that were judged as either entailment holds/no entailment by both judges. Yet, our evaluation target is to assess rules, and the Kappa values for the final correctness judgments of the shared rules were 0.74 and 0.68 for the lower and upper bound evaluations. These Kappa scores are regarded as 'substantial agreement' and are substantially higher than published agreement scores and those we managed to obtain using the standard rule-based approach. As expected, the agreement on rules is higher than on examples, since judges may disagree on a certain example but their judgements would still yield the same rule assessment.

Table 3 illustrates some disagreements that were still exhibited within the instance-based evaluation. The primary reason for disagreements was the difficulty to decide whether a context is relevant for a rule or not, resulting in some confusion between 'Irrelevant context' and 'No entailment'. This may explain the lower agreement for the upper bound precision, for which examples judged as 'Irrelevant context' are ignored, while for the lower bound both

[1] Available at http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool

[2] http://about.reuters.com/researchandstandards/corpus/

[3] Following advice by Patrick Pantel, DIRT's co-author.

Rule	Sentence	Judge 1	Judge 2
X sign $Y \rightarrow X$ set Y	**Iraq and Turkey** sign **agreement** to increase trade cooperation	Entailment holds	Irrelevant context
X worsen $Y \rightarrow X$ slow Y	**News of the strike** worsened **the situation**	Irrelevant context	No entailment
X get $Y \rightarrow X$ want Y	**He** will get **his parade** on Tuesday	Entailment holds	No entailment

Table 3: Examples for disagreement between the two judges.

judgments are conflated and represent no entailment. Our findings suggest that better ways for distinguishing relevant contexts may be sought in future research for further refinement of the instance-based evaluation methodology.

About 43% of all examples were judged as 'Left not entailed'. The relatively low matching precision (57%) made us collect more examples than needed, since 'Left not entailed' examples are ignored. Better matching capabilities will allow collecting and judging fewer examples, thus improving the efficiency of the evaluation process.

6 DIRT and TEASE Evaluation Results

	DIRT		TEASE	
	P	Y	P	Y
Rules:				
Upper Bound	30.5%	33.5	28.4%	40.3
Lower Bound	18.6%	20.4	17%	24.1
Templates:				
Upper Bound	44%	22.6	38%	26.9
Lower Bound	27.3%	14.1	23.6%	16.8

Table 4: Average Precision (P) and Yield (Y) at the rule and template levels.

We evaluated the quality of the entailment rules produced by each algorithm using two scores: (1) micro average ***Precision***, the percentage of correct rules out of all learned rules, and (2) average ***Yield***, the average number of correct rules learned for each input template I, as extrapolated based on the sample[4]. Since DIRT and TEASE do not identify rule directionality, we also measured these scores at the template level, where an output template O is considered correct if at least one of the rules '$I \rightarrow O$' or '$O \rightarrow I$' is correct. The results are presented in Table 4. The major finding is that the overall quality of DIRT and TEASE is very similar. Under the specific DIRT cutoff threshold chosen, DIRT exhibits somewhat higher Precision while TEASE has somewhat higher Yield (recall that there is no particular natural cutoff point for DIRT's output).

Since applications typically apply rules in a specific direction, the Precision for rules reflects their expected performance better than the Precision for templates. Obviously, future improvement in precision is needed for rule learning algorithms. Meanwhile, manual filtering of the learned rules can prove effective within limited domains, where our evaluation approach can be utilized for reliable filtering as well. The substantial yield obtained by these algorithms suggest that they are indeed likely to be valuable for recall increase in semantic applications.

In addition, we found that only about 15% of the correct templates were learned by both algorithms, which implies that the two algorithms largely complement each other in terms of coverage. One explanation may be that DIRT is focused on the domain of the local corpus used (news articles for the published DIRT knowledge-base), whereas TEASE learns from the Web, extracting rules from multiple domains. Since Precision is comparable it may be best to use both algorithms in tandem.

We also measured whether O is a paraphrase of I, i.e. whether both '$I \rightarrow O$' and '$O \rightarrow I$' are correct. Only 20-25% of all correct templates were assessed as paraphrases. This stresses the significance of evaluating directional rules rather than only paraphrases. Furthermore, it shows that in order to improve precision, acquisition algorithms must identify rule directionality.

[4]Since the rules are matched against the full corpus (as in IR evaluations), it is difficult to evaluate their true recall.

About 28% of all 'Left entailed' examples were evaluated as 'Irrelevant context', yielding the large difference in precision between the upper and lower precision bounds. This result shows that in order to get closer to the upper bound precision, learning algorithms and applications need to identify the relevant contexts in which a rule should be applied.

Last, we note that the instance-based quality assessment corresponds to the corpus from which the example sentences were taken. It is therefore best to evaluate the rules using a corpus of the same domain from which they were learned, or the target application domain for which the rules will be applied.

7 Conclusions

Accurate learning of inference knowledge, such as entailment rules, has become critical for further progress of applied semantic systems. However, evaluation of such knowledge has been problematic, hindering further developments. The instance-based evaluation approach proposed in this paper obtained acceptable agreement levels, which are substantially higher than those obtained for the common rule-based approach.

We also conducted the first comparison between two state-of-the-art acquisition algorithms, DIRT and TEASE, using the new methodology. We found that their quality is comparable but they effectively complement each other in terms of rule coverage. Also, we found that most learned rules are not paraphrases but rather one-directional entailment rules, and that many of the rules are context sensitive. These findings suggest interesting directions for future research, in particular learning rule directionality and relevant contexts, issues that were hardly explored till now. Such developments can be then evaluated by the instance-based methodology, which was designed to capture these two important aspects of entailment rules.

Acknowledgements

The authors would like to thank Ephi Sachs and Iddo Greental for their evaluation. This work was partially supported by ISF grant 1095/05, the IST Programme of the European Community under the PASCAL Network of Excellence IST-2002-506778, and the ITC-irst/University of Haifa collaboration.

References

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Second PASCAL Challenge Workshop for Recognizing Textual Entailment*.

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

Gennaro Chierchia and Sally McConnell-Ginet. 2000. *Meaning and Grammar (2nd ed.): an introduction to semantics*. MIT Press, Cambridge, MA.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. *Lecture Notes in Computer Science*, 3944:177–190.

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

Dekang Lin. 1998. Dependency-based evaluation of minipar. In *Proceedings of the Workshop on Evaluation of Parsing Systems at LREC*.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of HLT-NAACL*.

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

Lorenza Romano, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli. 2006. Investigating a generic paraphrase-based approach for relation extraction. In *Proceedings of EACL*.

Satoshi Sekine. 2005. Automatic paraphrase discovery based on context and keywords between ne pairs. In *Proceedings of IWP*.

Yusuke Shinyama, Satoshi Sekine, Kiyoshi Sudo, and Ralph Grishman. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of HLT*.

Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. 2003. An improved extraction pattern representation model for automatic IE pattern acquisition. In *Proceedings of ACL*.

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