"Ask not what Textual Entailment can do for You..."

Mark Sammons **V.G.Vinod Vydiswaran** **Dan Roth**
University of Illinois at Urbana-Champaign
{mssammon|vgvinodv|danr}@illinois.edu

Abstract

We challenge the NLP community to participate in a large-scale, distributed effort to design and build resources for developing and evaluating solutions to new and existing NLP tasks in the context of Recognizing Textual Entailment. We argue that the single global label with which RTE examples are annotated is insufficient to effectively evaluate RTE system performance; to promote research on smaller, related NLP tasks, we believe more detailed annotation and evaluation are needed, and that this effort will benefit not just RTE researchers, but the NLP community as a whole. We use insights from successful RTE systems to propose a model for identifying and annotating textual inference phenomena in textual entailment examples. We present the results of a pilot annotation study that show this model is feasible and the results immediately useful.

1 Introduction

Much of the work in the field of Natural Language Processing is founded on an assumption of semantic compositionality: that there are identifiable, separable components of an unspecified inference process that will develop as research in NLP progresses. Tasks such as Named Entity and coreference resolution, syntactic and shallow semantic parsing, and information and relation extraction have been identified as worthwhile tasks and pursued by numerous researchers. While many have (nearly) immediate application to real world tasks like search, many are also motivated by their potential contribution to more ambitious Natural Language tasks. It is clear that the components/tasks identified so far do not suffice in themselves to solve tasks requiring more complex reasoning and synthesis of information; many other tasks must be solved to achieve human-like performance on tasks such as Question Answering. But there is no clear process for identifying potential tasks (other than consensus by a sufficient number of researchers), nor for quantifying their potential contribution to existing NLP tasks, let alone to Natural Language Understanding.

Recent "grand challenges" such as *Learning by Reading*, *Learning To Read*, and *Machine Reading* are prompting more careful thought about the way these tasks relate, and what tasks must be solved in order to understand text sufficiently well to reliably reason with it. This is an appropriate time to consider a **systematic process for identifying semantic analysis tasks relevant to natural language understanding, and for assessing their potential impact on NLU system performance**.

Research on Recognizing Textual Entailment (RTE), largely motivated by a "grand challenge" now in its sixth year, has already begun to address some of the problems identified above. Techniques developed for RTE have now been successfully applied in the domains of Question Answering (Harabagiu and Hickl, 2006) and Machine Translation (Pado et al., 2009), (Mirkin et al., 2009). The RTE challenge examples are drawn from multiple domains, providing a relatively task-neutral setting in which to evaluate contributions of different component solutions, and RTE researchers have already made incremental progress by identifying sub-problems of entailment, and developing ad-hoc solutions for them.

In this paper we challenge the NLP community to contribute to a joint, long-term effort to identify, formalize, and solve textual inference problems motivated by the Recognizing Textual Entailment setting, in the following ways:

(a) Making the Recognizing Textual Entailment setting a central component of evaluation for

relevant NLP tasks such as NER, Coreference, parsing, data acquisition and application, and others. While many "component" tasks are considered (almost) solved in terms of expected improvements in performance on task-specific corpora, it is not clear that this translates to strong performance in the RTE domain, due either to problems arising from unrelated, unsolved entailment phenomena that co-occur in the same examples, or to domain change effects. The RTE task offers an application-driven setting for evaluating a broad range of NLP solutions, and will reinforce good practices by NLP researchers. The RTE task has been designed specifically to exercise textual inference capabilities, in a format that would make RTE systems potentially useful components in other "deep" NLP tasks such as Question Answering and Machine Translation. [1]

(b) Identifying relevant linguistic phenomena, interactions between phenomena, and their likely impact on RTE/textual inference. Determining the correct label for a single textual entailment example requires human analysts to make many smaller, localized decisions which may depend on each other. A broad, carefully conducted effort to identify and annotate such local phenomena in RTE corpora would allow their distributions in RTE examples to be quantified, and allow evaluation of NLP solutions in the context of RTE. It would also allow assessment of the potential impact of a solution to a specific sub-problem on the RTE task, and of interactions between phenomena. Such phenomena will almost certainly correspond to elements of linguistic theory; but this approach brings a data-driven approach to focus attention on those phenomena that are well-represented in the RTE corpora, and which can be identified with sufficiently close agreement.

(c) Developing resources and approaches that allow more detailed assessment of RTE systems. At present, it is hard to know what specific capabilities different RTE systems have, and hence, which aspects of successful systems are worth emulating or reusing. An evaluation framework that could offer insights into the kinds of sub-problems a given system can reliably solve would make it easier to identify significant advances, and thereby promote more rapid advances through reuse of successful solutions and focus on unresolved problems.

In this paper we demonstrate that Textual Entailment systems are already "interesting", in that they have made significant progress beyond a "smart" lexical baseline that is surprisingly hard to beat (section 2). We argue that Textual Entailment, as an application that clearly requires sophisticated textual inference to perform well, requires the solution of a range of sub-problems, some familiar and some not yet known. We therefore propose RTE as a promising and worthwhile task for large-scale community involvement, as it motivates the study of many other NLP problems in the context of general textual inference.

We outline the limitations of the present model of evaluation of RTE performance, and identify kinds of evaluation that would promote understanding of the way individual components can impact Textual Entailment system performance, and allow better objective evaluation of RTE system behavior without imposing additional burdens on RTE participants. We use this to motivate a large-scale annotation effort to provide data with the mark-up sufficient to support these goals.

To stimulate discussion of suitable annotation and evaluation models, we propose a candidate model, and provide results from a pilot annotation effort (section 3). This pilot study establishes the feasibility of an inference-motivated annotation effort, and its results offer a quantitative insight into the difficulty of the TE task, and the distribution of a number of entailment-relevant linguistic phenomena over a representative sample from the NIST TAC RTE 5 challenge corpus. We argue that such an evaluation and annotation effort can identify relevant subproblems whose solution will benefit not only Textual Entailment but a range of other long-standing NLP tasks, and can stimulate development of new ones. We also show how this data can be used to investigate the behavior of some of the highest-scoring RTE systems from the most recent challenge (section 4).

2 NLP Insights from Textual Entailment

The task of Recognizing Textual Entailment (RTE), as formulated by (Dagan et al., 2006), requires automated systems to identify when a human reader would judge that given one span of text (the Text) and some unspecified (but restricted) world knowledge, a second span of text (the Hy-

[1] The *Parser Training and Evaluation using Textual Entailment* track of SemEval 2 takes this idea one step further, by evaluating performance of an isolated NLP task using the RTE methodology.

Text:	The purchase of LexCorp by BMI for \$2Bn prompted widespread sell-offs by traders as they sought to minimize exposure.
Hyp 1:	BMI acquired another company.
Hyp 2:	BMI bought LexCorp for \$3.4Bn.

Figure 1: **Some representative RTE examples.**

pothesis) is true. The task was extended in (Giampiccolo et al., 2007) to include the additional requirement that systems identify when the Hypothesis contradicts the Text. In the example shown in figure 1, this means recognizing that the Text entails Hypothesis 1, while Hypothesis 2 contradicts the Text. This operational definition of Textual Entailment avoids commitment to any specific knowledge representation, inference method, or learning approach, thus encouraging application of a wide range of techniques to the problem.

2.1 An Illustrative Example

The simple RTE examples in figure 1 (most RTE examples have much longer Texts) illustrate some typical inference capabilities demonstrated by human readers in determining whether one span of text contains the meaning of another.

To recognize that Hypothesis 1 is entailed by the text, a human reader must recognize that “another company” in the Hypothesis can match “LexCorp”. She must also identify the nominalized relation “purchase”, and determine that “A purchased by B” implies “B acquires A”.

To recognize that Hypothesis 2 contradicts the Text, similar steps are required, together with the inference that because the stated purchase price is different in the Text and Hypothesis, but with high probability refers to the same transaction, Hypothesis 2 contradicts the Text.

It could be argued that this particular example might be resolved by simple lexical matching; but it should be evident that the Text can be made lexically very dissimilar to Hypothesis 1 while maintaining the Entailment relation, and that conversely, the lexical overlap between the Text and Hypothesis 2 can be made very high, while maintaining the Contradiction relation. This intuition is borne out by the results of the RTE challenges, which show that lexical similarity-based systems are outperformed by systems that use other, more structured analysis, as shown in the next section.

Rank	**System id**	**Accuracy**
1	I	0.735
2	E	0.685
3	H	0.670
4	J	0.667
5	G	0.662
6	B	0.638
7	D	0.633
8	F	0.632
9	A	0.615
9	C	0.615
9	K	0.615
-	Lex	0.612

Table 1: Top performing systems in the RTE 5 2-way task.

	Lex	**E**	**G**	**H**	**I**	**J**
Lex	**1.000** (184,183)	**0.667** (157,132)	**0.693** (168,122)	**0.678** (152,136)	**0.660** (165,137)	**0.778** (165,135)
E		**1.000** (224,187)	**0.667** (192,112)	**0.675** (178,131)	**0.673** (201,127)	**0.702** (186,131)
G			**1.000** (247,150)	**0.688** (186,120)	**0.713** (218,115)	**0.745** (198,125)
H				**1.000** (219,183)	**0.705** (194,139)	**0.707** (178,136)
I					**1.000** (260,181)	**0.705** (198,135)
J						**1.000** (224,178)

Table 2: In each cell, top row shows observed agreement and bottom row shows the number of correct (positive, negative) examples on which the pair of systems agree.

2.2 The State of the Art in RTE 5

The outputs for all systems that participated in the RTE 5 challenge were made available to participants. We compared these to each other and to a smart lexical baseline (Do et al., 2010) (lexical match augmented with a WordNet similarity measure, stemming, and a large set of low-semantic-content stopwords) to assess the diversity of the approaches of different research groups. To get the fullest range of participants, we used results from the two-way RTE task. We have anonymized the system names.

Table 1 shows that many participating systems significantly outperform our smart lexical baseline. Table 2 reports the observed agreement between systems and the lexical baseline in terms of the percentage of examples on which a pair of systems gave the same label. The agreement between most systems and the baseline is about 67%, which suggests that systems are not simply augmented versions of the lexical baseline, and are also distinct from each other in their behaviors.[2]

Common characteristics of RTE systems re-

[2] Note that the expected agreement between two random RTE decision-makers is 0.5, so the agreement scores according to Cohen’s Kappa measure (Cohen, 1960) are between 0.3 and 0.4.

ported by their designers were the use of structured representations of shallow semantic content (such as augmented dependency parse trees and semantic role labels); the application of NLP resources such as Named Entity recognizers, syntactic and dependency parsers, and coreference resolvers; and the use of special-purpose ad-hoc modules designed to address specific entailment phenomena the researchers had identified, such as the need for numeric reasoning. However, it is not possible to objectively assess the role these capabilities play in each system's performance from the system outputs alone.

2.3 The Need for Detailed Evaluation

An ablation study that formed part of the official RTE 5 evaluation attempted to evaluate the contribution of publicly available knowledge resources such as WordNet (Fellbaum, 1998), VerbOcean (Chklovski and Pantel, 2004), and DIRT (Lin and Pantel, 2001) used by many of the systems. The observed contribution was in most cases limited or non-existent. It is premature, however, to conclude that these resources have little potential impact on RTE system performance: most RTE researchers agree that the real contribution of individual resources is difficult to assess. As the example in figure 1 illustrates, most RTE examples require a number of phenomena to be correctly resolved in order to reliably determine the correct label (the Interaction problem); a perfect coreference resolver might as a result yield little improvement on the standard RTE evaluation, even though coreference resolution is clearly required by human readers in a significant percentage of RTE examples.

Various efforts have been made by individual research teams to address specific capabilities that are intuitively required for good RTE performance, such as (de Marneffe et al., 2008), and the formal treatment of entailment phenomena in (MacCartney and Manning, 2009) depends on and formalizes a divide-and-conquer approach to entailment resolution. But the phenomena-specific capabilities described in these approaches are far from complete, and many are not yet invented. To devote real effort to identify and develop such capabilities, researchers must be confident that the resources (and the will!) exist to create and evaluate their solutions, and that the resource can be shown to be relevant to a sufficiently large subset of the NLP community. While there is widespread belief that there are many relevant entailment phenomena, though each individually may be relevant to relatively few RTE examples (the Sparseness problem), we know of no systematic analysis to determine what those phenomena are, and how sparsely represented they are in existing RTE data.

If it were even known what phenomena were relevant to specific entailment examples, it might be possible to more accurately distinguish system capabilities, and promote adoption of successful solutions to sub-problems. An annotation-side solution also maintains the desirable agnosticism of the RTE problem formulation, by not imposing the requirement on system developers of generating an explanation for each answer. Of course, if examples were also annotated with explanations in a consistent format, this could form the basis of a new evaluation of the kind essayed in the pilot study in (Giampiccolo et al., 2007).

3 Annotation Proposal and Pilot Study

As part of our challenge to the NLP community, we propose a distributed OntoNotes-style approach (Hovy et al., 2006) to this annotation effort: distributed, because it should be undertaken by a diverse range of researchers with interests in different semantic phenomena; and similar to the OntoNotes annotation effort because it should not presuppose a fixed, closed ontology of entailment phenomena, but rather, iteratively hypothesize and refine such an ontology using inter-annotator agreement as a guiding principle. Such an effort would require a steady output of RTE examples to form the underpinning of these annotations; and in order to get sufficient data to represent less common, but nonetheless important, phenomena, a large body of data is ultimately needed.

A research team interested in annotating a new phenomenon should use examples drawn from the common corpus. Aside from any task-specific gold standard annotation they add to the entailment pairs, they should augment existing explanations by indicating in which examples their phenomenon occurs, and at which point in the existing explanation for each example. In fact, this latter effort – identifying phenomena relevant to textual inference, marking relevant RTE examples, and generating explanations – itself enables other researchers to select from known problems, assess their likely impact, and automatically generate relevant corpora.

To assess the feasibility of annotating RTE-oriented local entailment phenomena, we developed an inference model that could be followed by annotators, and conducted a pilot annotation study. We based our initial effort on observations about RTE data we made while participating in RTE challenges, together with intuitive conceptions of the kinds of knowledge that might be available in semi-structured or structured form. In this section, we present our annotation inference model, and the results of our pilot annotation effort.

3.1 Inference Process

To identify and annotate RTE sub-phenomena in RTE examples, we need a defensible model for the entailment process that will lead to consistent annotation by different researchers, and to an extensible framework that can accommodate new phenomena as they are identified.

We modeled the entailment process as one of manipulating the text and hypothesis to be as similar as possible, by first identifying parts of the text that matched parts of the hypothesis, and then identifying connecting structure. Our inherent assumption was that the meanings of the Text and Hypothesis could be represented as sets of n-ary relations, where relations could be connected to other relations (i.e., could take other relations as arguments). As we followed this procedure for a given example, we marked which entailment phenomena were required for the inference. We illustrate the process using the example in figure 1.

First, we would identify the arguments "BMI" and "another company" in the Hypothesis as matching "BMI" and "LexCorp" respectively, requiring 1) *Parent-Sibling* to recognize that "LexCorp" can match "company". We would tag the example as requiring 2) *Nominalization Resolution* to make "purchase" the active relation and 3) *Passivization* to move "BMI" to the subject position. We would then tag it with 4) *Simple Verb Rule* to map "A purchase B" to "A acquire B". These operations make the relevant portion of the Text identical to the Hypothesis, so we are done.

For the same Text, but with Hypothesis 2 (a negative example), we follow the same steps 1-3. We would then use 4) *Lexical Relation* to map "purchase" to "buy". We would then observe that the only possible match for the hypothesis argument "for $3.4Bn" is the text argument "for $2Bn". We would label this as a 5) *Numerical Quantity Mismatch* and 6) *Excluding Argument* (it can't be the case that in the same transaction, the same company was sold for two different prices).

Note that neither explanation mentions the anaphora resolution connecting "they" to "traders", because it is not strictly required to determine the entailment label.

As our example illustrates, this process makes sense for both positive and negative examples. It also reflects common approaches in RTE systems, many of which have explicit alignment components that map parts of the Hypothesis to parts of the Text prior to a final decision stage.

3.2 Annotation Labels

We sought to identify roles for background knowledge in terms of domains and general inference steps, and the types of linguistic phenomena that are involved in representing the same information in different ways, or in detecting key differences in two similar spans of text that indicate a difference in meaning. We annotated examples with domains (such as "Work") for two reasons: to establish whether some phenomena are correlated with particular domains; and to identify domains that are sufficiently well-represented that a knowledge engineering study might be possible.

While we did not generate an explicit representation of our entailment process, i.e. explanations, we tracked which phenomena were strictly required for inference. The annotated corpora and simple CGI scripts for annotation are available at *http://cogcomp.cs.illinois.edu/Data/ACL2010_RTE.php*.

The phenomena that we considered during annotation are presented in Tables 3, 4, 5, and 6. We tried to define each phenomenon so that it would apply to both positive and negative examples, but ran into a problem: often, negative examples can be identified principally by structural differences: the components of the Hypothesis all match components in the Text, but they are not connected by the appropriate structure in the Text. In the case of contradictions, it is often the case that a key relation in the Hypothesis must be matched to an incompatible relation in the Text. We selected names for these structural behaviors, and tagged them when we observed them, but the counterpart for positive examples must always hold: it must necessarily be the case that the structure in the Text linking the arguments that match those in the

Hypothesis must be comparable to the Hypothesis structure. We therefore did not tag this for positive examples.

We selected a subset of 210 examples from the NIST TAC RTE 5 (Bentivogli et al., 2009) Test set drawn equally from the three sub-tasks (IE, IR and QA). Each example was tagged by both annotators. Two passes were made over the data: the first covered 50 examples from each RTE sub-task, while the second covered an additional 20 examples from each sub-task. Between the two passes, concepts the annotators identified as difficult to annotate were discussed and more carefully specified, and several new concepts were introduced based on annotator observations.

Tables 3, 4, 5, and 6 present information about the distribution of the phenomena we tagged, and the inter-annotator agreement (Cohen's Kappa (Cohen, 1960)) for each. "Occurrence" lists the average percentage of examples labeled with a phenomenon by the two annotators.

Domain	Occurrence	Agreement
work	16.90%	0.918
name	12.38%	0.833
die kill injure	12.14%	0.979
group	9.52%	0.794
be in	8.57%	0.888
kinship	7.14%	1.000
create	6.19%	1.000
cause	6.19%	0.854
come from	5.48%	0.879
win compete	3.10%	0.813
Others	29.52%	0.864

Table 3: Occurrence statistics for domains in the annotated data.

Phenomenon	Occurrence	Agreement
Named Entity	91.67%	0.856
locative	17.62%	0.623
Numerical Quantity	14.05%	0.905
temporal	5.48%	0.960
nominalization	4.05%	0.245
implicit relation	1.90%	0.651

Table 4: Occurrence statistics for hypothesis structure features.

From the tables it is apparent that good performance on a range of phenomena in our inference model are likely to have a significant effect on RTE results, with coreference being deemed essential to the inference process for 35% of examples, and a number of other phenomena are sufficiently well represented to merit near-future attention (assuming that RTE systems do not already handle these phenomena, a question we address in section 4). It is also clear from the predominance of *Simple Rewrite Rule* instances, together with the frequency of most of the domains we selected, that knowledge engineering efforts also have a key role in improving RTE performance.

Phenomenon	Occurrence	Agreement
coreference	35.00%	0.698
simple rewrite rule	32.62%	0.580
lexical relation	25.00%	0.738
implicit relation	23.33%	0.633
factoid	15.00%	0.412
parent-sibling	11.67%	0.500
genetive relation	9.29%	0.608
nominalization	8.33%	0.514
event chain	6.67%	0.589
coerced relation	6.43%	0.540
passive-active	5.24%	0.583
numeric reasoning	4.05%	0.847
spatial reasoning	3.57%	0.720

Table 5: Occurrence statistics for entailment phenomena and knowledge resources

Phenomenon	Occurrence	Agreement
missing argument	16.19%	0.763
missing relation	14.76%	0.708
excluding argument	10.48%	0.952
Named Entity mismatch	9.29%	0.921
excluding relation	5.00%	0.870
disconnected relation	4.52%	0.580
missing modifier	3.81%	0.465
disconnected argument	3.33%	0.764
Numeric Quant. mismatch	3.33%	0.882

Table 6: Occurrences of negative-only phenomena

3.3 Discussion

Perhaps surprisingly, given the difficulty of the task, inter-annotator agreement was consistently good to excellent (above 0.6 and 0.8, respectively), with few exceptions, indicating that for most targeted phenomena, the concepts were well-specified. The results confirmed our initial intuition about some phenomena: for example, that coreference resolution is central to RTE, and that detecting the connecting structure is crucial in discerning negative from positive examples. We also found strong evidence that the difference between contradiction and unknown entailment examples is often due to the behavior of certain relations that either preclude certain other relations holding between the same arguments (for example, winning a contest vs. losing a contest), or which can only hold for a single referent in one argument position (for example, "work" relations such as job title are typically constrained so that a single person holds one position).

We found that for some examples, there was more than one way to infer the hypothesis from the text. Typically, for positive examples this involved overlap between phenomena; for example, Coreference might be expected to resolve implicit relations induced from appositive structures. In such cases we annotated every way we could find.

In future efforts, annotators should record the entailment steps they used to reach their decision. This will make disagreement resolution simpler, and could also form a possible basis for generating gold standard explanations. At a minimum, each inference step must identify the spans of the Text and Hypothesis that are involved and the name of the entailment phenomenon represented; in addition, a partial order over steps must be specified when one inference step requires that another has been completed.

Future annotation efforts should also add a category "Other", to indicate for each example whether the annotator considers the listed entailment phenomena sufficient to identify the label. It might also be useful to assess the difficulty of each example based on the time required by the annotator to determine an explanation, for comparison with RTE system errors.

These, together with specifications that minimize the likely disagreements between different groups of annotators, are processes that must be refined as part of the broad community effort we seek to stimulate.

4 Pilot RTE System Analysis

In this section, we sketch out ways in which the proposed analysis can be applied to learn something about RTE system behavior, even when those systems do not provide anything beyond the output label. We present the analysis in terms of sample questions we hope to answer with such an analysis.

1. If a system needs to improve its performance, which features should it concentrate on? To answer this question, we looked at the top-5 systems and tried to find which phenomena are active in the mistakes they make.

(a) Most systems seem to fail on examples that need *numeric reasoning* to get the entailment decision right. For example, system H got all 10 examples with numeric reasoning wrong.

(b) All top-5 systems make consistent errors in cases where identifying a mismatch in named entities (NE) or numerical quantities (NQ) is important to make the right decision. System G got 69% of cases with NE/NQ mismatches wrong.

(c) Most systems make errors in examples that have a disconnected or exclusion component (argument/relation). System J got 81% of cases with a disconnected component wrong.

(d) Some phenomena are handled well by certain systems, but not by others. For example, failing to recognize a parent-sibling relation between entities/concepts seems to be one of the top-5 phenomena active in systems E and H. System H also fails to correctly label over 53% of the examples having kinship relation.

2. Which phenomena have strong correlations to the entailment labels among hard examples? We called an example hard if at least 4 of the top 5 systems got the example wrong. In our annotation dataset, there were 41 hard examples. Some of the phenomena that strongly correlate with the TE labels on hard examples are: deeper lexical relation between words ($\rho = 0.542$), and need for external knowledge ($\rho = 0.345$). Further, we find that the top-5 systems tend to make mistakes in cases where the lexical approach also makes mistakes ($\rho = 0.355$).

3. What more can be said about individual systems? In order to better understand the system behavior, we wanted to check if we could predict the system behavior based on the phenomena we identified as important in the examples. We learned SVM classifiers over the identified phenomena and the lexical similarity score to predict both the labels and errors systems make for each of the top-5 systems. We could predict all 10 system behaviors with over 70% accuracy, and could predict labels and mistakes made by two of the top-5 systems with over 77% accuracy. This indicates that although the identified phenomena are indicative of the system performance, it is probably too simplistic to assume that system behavior can be easily reproduced solely as a disjunction of phenomena present in the examples.

4. Does identifying the phenomena correctly help learn a better TE system? We tried to learn an entailment classifier over the phenomenon identified and the top 5 system outputs. The results are summarized in Table 7. All reported numbers are 20-fold cross-validation accuracy from an SVM classifier learned over the features mentioned. The results show that correctly identifying the named-entity and numeric quantity mis-

No.	Feature description	No. of	Accuracy over which features	
		feats	phenomena	pheno. + sys. labels
(0)	Only system labels	5	—	0.714
(1)	Domain and hypothesis features (Tables 3, 4)	16	0.510	0.705
(2)	(1) + NE + NQ	18	0.619	0.762
(3)	(1) + Knowledge resources (subset of Table 5)	22	0.662	0.762
(4)	(3) + NE + NQ	24	0.738	0.805
(5)	(1) + Entailment and Knowledge resources (Table 5)	29	0.748	0.791
(6)	(5) + negative-only phenomena (Table 6)	38	0.971	0.943

Table 7: Accuracy in predicting the label based on the phenomena and top-5 system labels.

matches improves the overall accuracy significantly. If we further recognize the need for knowledge resources correctly, we can correctly explain the label for 80% of the examples. Adding the entailment and negation features helps us explain the label for 97% of the examples in the annotated corpus.

It must be clarified that the results do not show the textual entailment problem itself is solved with 97% accuracy. However, we believe that if a system could recognize key negation phenomena such as *Named Entity mismatch*, presence of *Excluding arguments*, etc. correctly and consistently, it could model them as a *Contradiction* features in the final inference process to significantly improve its overall accuracy. Similarly, identifying and resolving the key entailment phenomena in the examples, would boost the inference process in positive examples. However, significant effort is still required to obtain near-accurate knowledge and linguistic resources.

5 Discussion

NLP researchers in the broader community continually seek new problems to solve, and pose more ambitious tasks to develop NLP and NLU capabilities, yet recognize that even solutions to problems which are considered "solved" may not perform as well on domains different from the resources used to train and develop them. Solutions to such NLP tasks could benefit from evaluation and further development on corpora drawn from a range of domains, like those used in RTE evaluations.

It is also worthwhile to consider each task as part of a larger inference process, and therefore motivated not just by performance statistics on special-purpose corpora, but as part of an interconnected web of resources; and the task of Recognizing Textual Entailment has been designed to exercise a wide range of linguistic and reasoning capabilities.

The entailment setting introduces a potentially broader context to resource development and assessment, as the hypothesis and text provide context for each other in a way different than local context from, say, the same paragraph in a document: in RTE's positive examples, the Hypothesis either restates some part of the Text, or makes statements inferable from the statements in the Text. This is not generally true of neighboring sentences in a document. This distinction opens the door to "purposeful", or goal-directed, inference in a way that may not be relevant to a task studied in isolation.

The RTE community seems mainly convinced that incremental advances in local entailment phenomena (including application of world knowledge) are needed to make significant progress. They need ways to identify sub-problems of textual inference, and to evaluate those solutions both in isolation and in the context of RTE. RTE system developers are likely to reward well-engineered solutions by adopting them and citing their authors, because such solutions are easier to incorporate into RTE systems. They are also more likely to adopt solutions with established performance levels. These characteristics promote publication of software developed to solve NLP tasks, attention to its usability, and publication of materials supporting reproduction of results presented in technical papers.

For these reasons, we assert that RTE is a natural motivator of new NLP tasks, as researchers look for components capable of improving performance; and that RTE is a natural setting for evaluating solutions to a broad range of NLP problems, though not in its present formulation: we must solve the problem of credit assignment, to recognize component contributions. We have therefore proposed a suitable annotation effort, to provide the resources necessary for more detailed evaluation of RTE systems.

We have presented a linguistically-motivated analysis of entailment data based on a step-wise procedure to resolve entailment decisions, intended to allow independent annotators to reach consistent decisions, and conducted a pilot annotation effort to assess the feasibility of such a task.

We do not claim that our set of domains or phenomena are complete: for example, our illustrative example could be tagged with a domain *Mergers and Acquisitions*, and a different team of researchers might consider *Nominalization Resolution* to be a subset of *Simple Verb Rules*. This kind of disagreement in coverage is inevitable, but we believe that in many cases it suffices to introduce a new domain or phenomenon, and indicate its relation (if any) to existing domains or phenomena. In the case of introducing a non-overlapping category, no additional information is needed. In other cases, the annotators can simply indicate the phenomena being merged or split (or even replaced). This information will allow other researchers to integrate different annotation sources and maintain a consistent set of annotations.

6 Conclusions

In this paper, we have presented a case for a broad, long-term effort by the NLP community to coordinate annotation efforts around RTE corpora, and to evaluate solutions to NLP tasks relating to textual inference in the context of RTE. We have identified limitations in the existing RTE evaluation scheme, proposed a more detailed evaluation to address these limitations, and sketched a process for generating this annotation. We have proposed an initial annotation scheme to prompt discussion, and through a pilot study, demonstrated that such annotation is both feasible and useful.

We ask that researchers not only contribute task specific annotation to the general pool, and indicate how their task relates to those already added to the annotated RTE corpora, but also invest the additional effort required to augment the cross-domain annotation: marking the examples in which their phenomenon occurs, and augmenting the annotator-generated explanations with the relevant inference steps.

These efforts will allow a more meaningful evaluation of RTE systems, and of the component NLP technologies they depend on. We see the potential for great synergy between different NLP subfields, and believe that all parties stand to gain from this collaborative effort. We therefore respectfully suggest that you “ask not what RTE can do for you, but what you can do for RTE...”

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. This research was partly sponsored by Air Force Research Laboratory (AFRL) under prime contract no. FA8750-09-C-0181, by a grant from Boeing, and by MIAS, the Multimodal Information Access and Synthesis center at UIUC, part of CCICADA, a DHS Center of Excellence. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the sponsors.

References

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernando Magnini. 2009. The fifth pascal recognizing textual entailment challenge. In *Notebook papers and Results, Text Analysis Conference (TAC)*, pages 14–24.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, pages 33–40.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

I. Dagan, O. Glickman, and B. Magnini, editors. 2006. *The PASCAL Recognising Textual Entailment Challenge.*, volume 3944. Springer-Verlag, Berlin.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio, June. Association for Computational Linguistics.

Quang Do, Dan Roth, Mark Sammons, Yuancheng Tu, and V.G.Vinod Vydiswaran. 2010. Robust, Light-weight Approaches to compute Lexical Similarity. Computer Science Research and Technical Reports, University of Illinois. http://L2R.cs.uiuc.edu/∼danr/Papers/DRSTV10.pdf.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague, June. Association for Computational Linguistics.

Sanda Harabagiu and Andrew Hickl. 2006. Methods for Using Textual Entailment in Open-Domain Question Answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 905–912, Sydney, Australia, July. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of HLT/NAACL*, New York.

D. Lin and P. Pantel. 2001. DIRT: discovery of inference rules from text. In *Proc. of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*, pages 323–328.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *The Eighth International Conference on Computational Semantics (IWCS-8)*, Tilburg, Netherlands.

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. Source-language entailment modeling for translating unknown terms. In *ACL/AFNLP*, pages 791–799, Suntec, Singapore, August. Association for Computational Linguistics.

Sebastian Pado, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 297–305, Suntec, Singapore, August. Association for Computational Linguistics.