Contradictions and Justifications: Extensions to the Textual Entailment Task

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, USA
ellen.voorhees@nist.gov

Abstract

The third PASCAL Recognizing Textual Entailment Challenge (RTE-3) contained an optional task that extended the main entailment task by requiring a system to make three-way entailment decisions (entails, contradicts, neither) and to justify its response. Contradiction was rare in the RTE-3 test set, occurring in only about 10% of the cases, and systems found accurately detecting it difficult. Subsequent analysis of the results shows a test set must contain many more entailment pairs for the three-way decision task than the traditional two-way task to have equal confidence in system comparisons. Each of six human judges representing eventual end users rated the quality of a justification by assigning "understandability" and "correctness" scores. Ratings of the same justification across judges differed significantly, signaling the need for a better characterization of the justification task.

1 Introduction

The PASCAL Recognizing Textual Entailment (RTE) workshop series (see www.pascal-network.org/Challenges/RTE3/) has been a catalyst for recent research in developing systems that are able to detect when the content of one piece of text necessarily follows from the content of another piece of text (Dagan et al., 2006; Giampiccolo et al., 2007). This ability is seen as a fundamental component in the solutions for a variety of natural language problems such as question answering, summarization, and information extraction. In addition to the main entailment task, the most recent Challenge, RTE-3, contained a second optional task that extended the main task in two ways. The first extension was to require systems to make three-way entailment decisions; the second extension was for systems to return a justification or explanation of how its decision was reached.

In the main RTE entailment task, systems report whether the *hypothesis* is entailed by the *text*. The system responds with YES if the hypothesis is entailed and NO otherwise. But this binary decision conflates the case when the hypothesis actually contradicts the text—the two could not both be true—with simple lack of entailment. The three-way entailment decision task requires systems to decide whether the hypothesis is entailed by the text (YES), contradicts the text (NO), or is neither entailed by nor contradicts the text (UNKNOWN).

The second extension required a system to explain why it reached its conclusion in terms suitable for an eventual end user (i.e., not system developer). Explanations are one way to build a user's trust in a system, but it is not known what kinds of information must be conveyed nor how best to present that information. RTE-3 provided an opportunity to collect a diverse sample of explanations to begin to explore these questions.

This paper analyzes the extended task results, with the next section describing the three-way decision subtask and Section 3 the justification subtask. Contradiction was rare in the RTE-3 test set, occurring in only about 10% of the cases, and systems found accurately detecting it difficult. While the level of agreement among human annotators as to

Proceedings of ACL-08: HLT, pages 63–71,
Columbus, Ohio, USA, June 2008.

the correct answer for an entailment pair was within expected bounds, the test set was found to be too small to reliably distinguish among systems' three-way accuracy scores. Human judgments of the quality of a justification varied widely, signaling the need for a better characterization of the justification task. Comments from the judges did include some common themes. Judges prized conciseness, though they were uncomfortable with mathematical notation unless they had a mathematical background. Judges strongly disliked being shown system internals such as scores reported by various components.

2 The Three-way Decision Task

The extended task used the RTE-3 main task test set of entailment pairs as its test set. This test set contains 800 text and hypothesis pairs, roughly evenly split between pairs for which the text entails the hypothesis (410 pairs) and pairs for which it does not (390 pairs), as defined by the reference answer key released by RTE organizers.

RTE uses an "ordinary understanding" principle for deciding entailment. The hypothesis is considered entailed by the text if a human reading the text would most likely conclude that the hypothesis were true, even if there could exist unusual circumstances that would invalidate the hypothesis. It is explicitly acknowledged that ordinary understanding depends on a common human understanding of language as well as common background knowledge. The extended task also used the ordinary understanding principle for deciding contradictions. The hypothesis and text were deemed to contradict if a human would most likely conclude that the text and hypothesis could not both be true.

The answer key for the three-way decision task was developed at the National Institute of Standards and Technology (NIST) using annotators who had experience as TREC and DUC assessors. NIST assessors annotated all 800 entailment pairs in the test set, with each pair independently annotated by two different assessors. The three-way answer key was formed by keeping exactly the same set of YES answers as in the two-way key (regardless of the NIST annotations) and having NIST staff adjudicate assessor differences on the remainder. This resulted in a three-way answer key containing 410 (51%) YES answers, 319 (40%) UNKNOWN answers, and 72 (9%) NO answers.

Reference Answer	Systems' Responses YES	UNKN	NO	Totals
YES	2449	2172	299	4920
UNKN	929	2345	542	3816
NO	348	415	101	864
Totals	3726	4932	942	9600

Table 1: Contingency table of responses over all 800 entailment pairs and all 12 runs.

2.1 System results

Eight different organizations participated in the three-way decision subtask submitting a total of 12 runs. A run consists of exactly one response of YES, NO, or UNKNOWN for each of the 800 test pairs. Runs were evaluated using accuracy, the percentage of system responses that match the reference answer.

Figure 1 shows both the overall accuracy of each of the runs (numbers running along the top of the graph) and the accuracy as conditioned on the reference answer (bars). The conditioned accuracy for YES answers, for example, is accuracy computed using just those test pairs for which YES is the reference answer. The runs are sorted by decreasing overall accuracy.

Systems were much more accurate in recognizing entailment than contradiction (black bars are greater than white bars). Since conditioned accuracy does not penalize for overgeneration of a response, the conditioned accuracy for UNKNOWN is excellent for those systems that used UNKNOWN as their default response. Run H never concluded that a pair was a contradiction, for example.

Table 1 gives another view of the relative difficulty of detecting contradiction. The table is a contingency table of the systems' responses versus the reference answer summed over all test pairs and all runs. A reference answer is represented as a row in the table and a system's response as a column. Since there are 800 pairs in the test set and 12 runs, there is a total of 9600 responses.

As a group the systems returned NO as a response 942 times, approximately 10% of the time. While 10% is a close match to the 9% of the test set for which NO is the reference answer, the systems detected contradictions for the wrong pairs: the table's

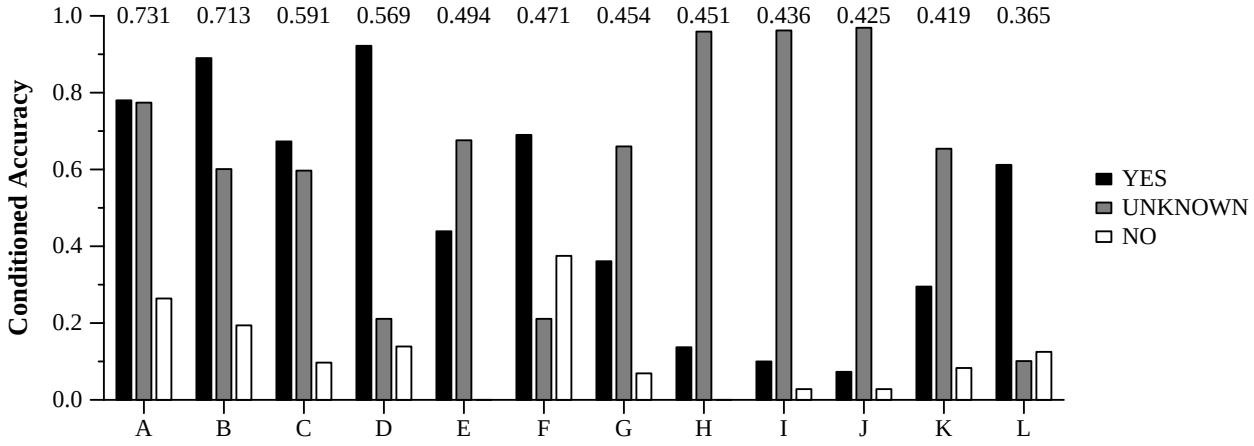

Figure 1: Overall accuracy (top number) and accuracy conditioned by reference answer for three-way runs.

diagonal entry for NO is the smallest entry in both its row and its column. The smallest row entry means that systems were more likely to respond that the hypothesis was entailed than that it contradicted when it in fact contradicted. The smallest column entry means than when the systems did respond that the hypothesis contradicted, it was more often the case that the hypothesis was actually entailed than that it contradicted. The 101 correct NO responses represent 12% of the 864 possible correct NO responses. In contrast, the systems responded correctly for 50% (2449/4920) of the cases when YES was the reference answer and for 61% (2345/3816) of the cases when UNKNOWN was the reference answer.

Main Task	NIST Judge 1		
	YES	UNKN	NO
YES	378	27	5
NO	48	242	100

conflated agreement = .90

Main Task	NIST Judge 2		
	YES	UNKN	NO
YES	383	23	4
NO	46	267	77

conflated agreement = .91

Table 2: Agreement between NIST judges (columns) and main task reference answers (rows).

2.2 Human agreement

Textual entailment is evaluated assuming that there is a single correct answer for each test pair. This is a simplifying assumption used to make the evaluation tractable, but as with most NLP phenomena it is not actually true. It is quite possible for two humans to have legitimate differences of opinions (i.e., to differ when neither is mistaken) about whether a hypothesis is entailed or contradicts, especially given annotations are based on ordinary understanding.

Since systems are given credit only when they respond with the reference answer, differences in annotators' opinions can clearly affect systems' accuracy scores. The RTE main task addressed this issue by including a candidate entailment pair in the test set only if multiple annotators agreed on its disposition (Giampiccolo et al., 2007). The test set also contains 800 pairs so an individual test case contributes only $1/800 = 0.00125$ to the overall accuracy score. To allow the results from the two- and three-way decision tasks to be comparable (and to leverage the cost of creating the main task test set), the extended task used the same test set as the main task and used simple accuracy as the evaluation measure. The expectation was that this would be as effective an evaluation design for the three-way task as it is for the two-way task. Unfortunately, subsequent analysis demonstrates that this is not so.

Recall that NIST judges annotated all 800 entailment pairs in the test set, with each pair independently annotated twice. For each entailment pair, one of the NIST judges was arbitrarily assigned as the first judge for that pair and the other as the second judge. The agreement between NIST and RTE annotators is shown in Table 2. The top half of

the table shows the agreement between the two-way answer key and the annotations of the set of first judges; the bottom half is the same except using the annotations of the set of second judges. The NIST judges' answers are given in the columns and the two-way reference answers in the rows. Each cell in the table gives the raw count before adjudication of the number of test cases that were assigned that combination of annotations. Agreement is then computed as the percentage of matches when a NIST judge's NO or UNKNOWN annotation matched a NO two-way reference answer. Agreement is essentially identical for both sets of judges at 0.90 and 0.91 respectively.

Because the agreement numbers reflect the raw counts before adjudication, at least some of the differences may be attributable to annotator errors that were corrected during adjudication. But there do exist legitimate differences of opinion, even for the extreme cases of entails versus contradicts. Typical disagreements involve granularity of place names and amount of background knowledge assumed. Example disagreements concerned whether Hollywood was equivalent to Los Angeles, whether East Jerusalem was equivalent to Jerusalem, and whether members of the same political party who were at odds with one another were 'opponents'.

RTE organizers reported an agreement rate of about 88% among their annotators for the two-way task (Giampiccolo et al., 2007). The 90% agreement rate between the NIST judges and the two-way answer key probably reflects a somewhat larger amount of disagreement since the test set already had RTE annotators' disagreements removed. But it is similar enough to support the claim that the NIST annotators agree with other annotators as often as can be expected. Table 3 shows the three-way agreement between the two NIST annotators. As above, the table gives the raw counts before adjudication and agreement is computed as percentage of matching annotations. Three-way agreement is 0.83—smaller than two-way agreement simply because there are more ways to disagree.

Just as annotator agreement declines as the set of possible answers grows, the inherent stability of the accuracy measure also declines: accuracy and agreement are both defined as the percentage of exact matches on answers. The increased uncertainty

	YES	UNKN	NO
YES	381		
UNKN	82	217	
NO	11	43	66

three-way agreement = .83

Table 3: Agreement between NIST judges.

when moving from two-way to three-way decisions significantly reduces the power of the evaluation. With the given level of annotator agreement and 800 pairs in the test set, in theory accuracy scores could change by as much as 136 (the number of test cases for which annotators disagreed) $\times 0.00125 = .17$ by using a different choice of annotator. The maximum difference in accuracy scores actually observed in the submitted runs was 0.063.

Previous analyses of other evaluation tasks such as document retrieval and question answering demonstrated that system rankings are stable despite differences of opinion in the underlying annotations (Voorhees, 2000; Voorhees and Tice, 2000). The differences in accuracy observed for the three-way task are large enough to affect system rankings, however. Compared to the system ranking of ABCDEFGHIJKL induced by the official three-way answer key, the ranking induced by the first set of judges' raw annotations is BADCFEGKHLIJ. The ranking induced by the second set of judges' raw annotations is much more similar to the official results, ABCDEFGHKIJL.

How then to proceed? Since the three-way decision task was motivated by the belief that distinguishing contradiction from simple non-entailment is important, reverting back to a binary decision task is not an attractive option. Increasing the size of the test set beyond 800 test cases will result in a more stable evaluation, though it is not known how big the test set needs to be. Defining new annotation rules in hopes of increasing annotator agreement is a satisfactory option only if those rules capture a characteristic of entailment that systems should actually embody. Reasonable people *do* disagree about entailment and it is unwise to enforce some arbitrary definition in the name of consistency. Using UNKNOWN as the reference answer for all entailment pairs on which annotators disagree may be a reasonable strategy: the disagreement itself is strong evidence that

neither of the other options holds. Creating balanced test sets using this rule could be difficult, however. Following this rule, the RTE-3 test set would have 360 (45%) YES answers, 64 (8%) NO answers, and 376 (47%) UNKNOWN answers, and would induce the ranking ABCDEHIJGKFL. (Runs such as H, I, and J that return UNKNOWN as a default response are rewarded using this annotation rule.)

3 Justifications

The second part of the extended task was for systems to provide explanations of how they reached their conclusions. The specification of a justification for the purposes of the task was deliberately vague—a collection of ASCII strings with no minimum or maximum size—so as to not preclude good ideas by arbitrary rules. A justification run contained all of the information from a three-way decision run plus the rationale explaining the response for each of the 800 test pairs in the RTE-3 test set. Six of the runs shown in Figure 1 (A, B, C, D, F, and H) are justification runs. Run A is a manual justification run, meaning there was some human tweaking of the justifications (but not the entailment decisions).

After the runs were submitted, NIST selected a subset of 100 test pairs to be used in the justification evaluation. The pairs were selected by NIST staff after looking at the justifications so as to maximize the informativeness of the evaluation set. All runs were evaluated on the same set of 100 pairs.

Figure 2 shows the justification produced by each run for pair 75 (runs D and F were submitted by the same organization and contained identical justifications for many pairs including pair 75). The text of pair 75 is *Muybridge had earlier developed an invention he called the Zoopraxiscope.*, and the hypothesis is *The Zoopraxiscope was invented by Muybridge.* The hypothesis is entailed by the text, and each of the systems correctly replied that it is entailed. Explanations for why the hypothesis is entailed differ widely, however, with some rationales of dubious validity.

Each of the six different NIST judges rated all 100 justifications. For a given justification, a judge first assigned an integer score between 1–5 on how understandable the justification was (with 1 as unintelligible and 5 as completely understandable). If the understandability score assigned was 3 or greater, the judge then assigned a correctness score, also an integer between 1–5 with 5 the high score. This second score was interpreted as how compelling the argument contained in the justification was rather than simple correctness because justifications could be strictly correct but immaterial.

3.1 System results

The motivation for the justification subtask was to gather data on how systems might best explain themselves to eventual end users. Given this goal and the exploratory nature of the exercise, judges were given minimal guidance on how to assign scores other than that it should be from a user's, not a system developer's, point of view. Judges used a system that displayed the text, hypothesis, and reference answer, and then displayed each submission's justification in turn. The order in which the runs' justifications were displayed was randomly selected for each pair; for a given pair, each judge saw the same order.

Figure 2 includes the scores assigned to each of the justifications of entailment pair 75. Each pair of numbers in brackets is a score pair assigned by one judge. The first number in the pair is the understandability score and the second the correctness score. The correctness score is omitted ('–') when the understandability score is 1 or 2 because no correctness score was assigned in that case. The scores from the different judges are given in the same order for each justification.

With 100 entailment pairs evaluated by each of 6 judges assigning 2 separate scores, each run had a total of 1200 numbers assigned to it. Figure 3 shows two views of these numbers: a histogram of the number of justifications in the run that were assigned a given score value summed over all judges and all test pairs, and the overall mean score for the run[1]. A correctness score that was not assigned because understandability was too poor is displayed as a score of 0 in the histogram and treated as a 0 in the computation of the mean. Understandability scores are shown toward the left in the figure and correctness scores toward the right.

The mean scores for correctness are fairly low

[1]Using the mean treats a score as an interval variable rather than an ordinal variable, a questionable treatment given a 5-point scale. A mean is a convenient summary, however.

Run A:

```
There is a relationship between Zoopraxiscope and Muybridge in both the text and
hypothesis.  The term "invention" is morphologically similar to "invented".
```

Scores: [4,4] [4,4] [5,4] [4,1] [5,4], [3,2]

Run B:

```
The text mentions 'Muybridge'.  We can infer that Muybridge is inventor.  From
the fact that Muybridge is inventor, we can infer that Muybridge invented.  We can
conclude that the Zoopraxiscope was invented by Muybridge.
```

Scores:[4,3] [3,3] [5,4] [5,1] [5,3] [3,2]

Run C:

```
1:  The Hypothesis could be precisely matched with content in the Text, with
allowance for polarity and embedded contexts.
2:  Hypothesis words match well with words in the Text.
3:  text adjunct "called" of "invention" dropped on aligned hyp word "invented"
```

Scores: [3,3] [4,4] [4,4] [2,–] [1,–] [2,–]

Runs D and F:

```
The words in the hypothesis are all found, with the approximately all the exact same
syntactic dependencies, also in the text.  Therefore, I concluded that the given
hypothesis:
   The Zoopraxiscope was invented by Muybridge.
is entailed by the given text
   Muybridge had earlier developed an invention he called the Zoopraxiscope.
```

Scores: [3,3] [4,3] [4,3] [5,1] [4,3] [2,–]

Run H:

```
Yes!
I have general knowledge that:
IF Y is developed by X THEN Y is manufactured by X
Here:  X = Muybridge, Y = the invention
Thus, here:
   We are told in T: the invention is developed by Muybridge
   Thus it follows that:  the invention is manufactured by Muybridge
In addition, I know:
   "manufacture" and "invent" mean roughly the same thing
Hence:  The Zoopraxiscope was invented by Muybridge.
```

Scores: [2,–] [4,1] [3,3] [3,1] [2,–] [1,–]

Figure 2: Justification for entailment pair 75 from each justification run. Brackets contain the pair of scores assigned to the justification by one of the six human judges; the first number in the pair is the understandability score and the second is the correctness score.

for all runs. Recall, however, that the 'correctness' score was actually interpreted as compellingness. There were many justifications that were strictly correct but not very informative, and they received low correctness scores. For example, the low correctness scores for the justification from run A in Figure 2 were given because those judges did not feel that the fact that "invention and inventor are morphologically similar" was enough of an explanation. Mean correctness scores were also affected by understandability. Since an unassigned correctness score was treated as a zero when computing the mean, systems with low understandability scores must have lower correctness scores. Nonetheless, it is also true that systems reached the correct entailment decision by faulty reasoning uncomfortably often, as illustrated by the justification from run H in Figure 2.

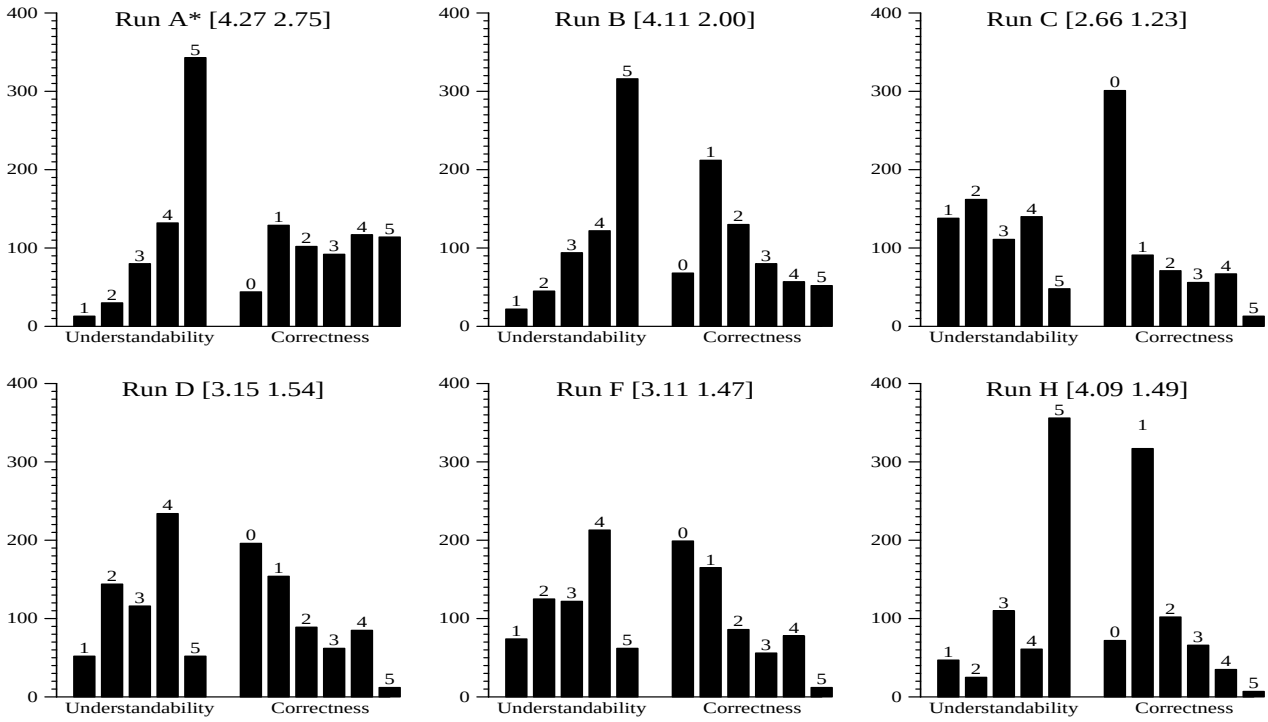

Figure 3: Number of justifications in a run that were assigned a particular score value summed over all judges and all test pairs. Brackets contain the overall mean understandability and correctness scores for the run. The starred run (A) is the manual run.

3.2 Human agreement

The most striking feature of the system results in Figure 3 is the variance in the scores. Not explicit in that figure, though illustrated in the example in Figure 2, is that different judges often gave widely different scores to the same justification. One systematic difference was immediately detected. The NIST judges have varying backgrounds with respect to mathematical training. Those with more training were more comfortable with, and often preferred, justifications expressed in mathematical notation; those with little training strongly disliked any mathematical notation in an explanation. This preference affected both the understandability and the correctness scores. Despite being asked to assign two separate scores, judges found it difficult to separate understandability and correctness. As a result, correctness scores were affected by presentation.

The scores assigned by different judges were sufficiently different to affect how runs compared to one another. This effect was quantified in the following way. For each entailment pair in the test set, the set of six runs was ranked by the scores assigned by one assessor, with rank one assigned to the best run and rank six the worst run. If several systems had the same score, they were each assigned the mean rank for the tied set. (For example, if two systems had the same score that would rank them second and third, they were each assigned rank 2.5.) A run was then assigned its mean rank over the 100 justifications. Figure 4 shows how the mean rank of the runs varies by assessor. The x-axis in the figure shows the judge assigning the score and the y-axis the mean rank (remember that rank one is best). A run is plotted using its letter name consistent with previous figures, and lines connect the same system across different judges. Lines intersect demonstrating that different judges prefer different justifications.

After rating the 100 justifications, judges were asked to write a short summary of their impression of the task and what they looked for in a justification. These summaries did have some common themes. Judges prized conciseness and specificity, and expected (or at least hoped for) explanations in fluent English. Judges found "chatty" templates such as the one used in run H more annoying than engaging. Verbatim repetition of the text and hypothesis within

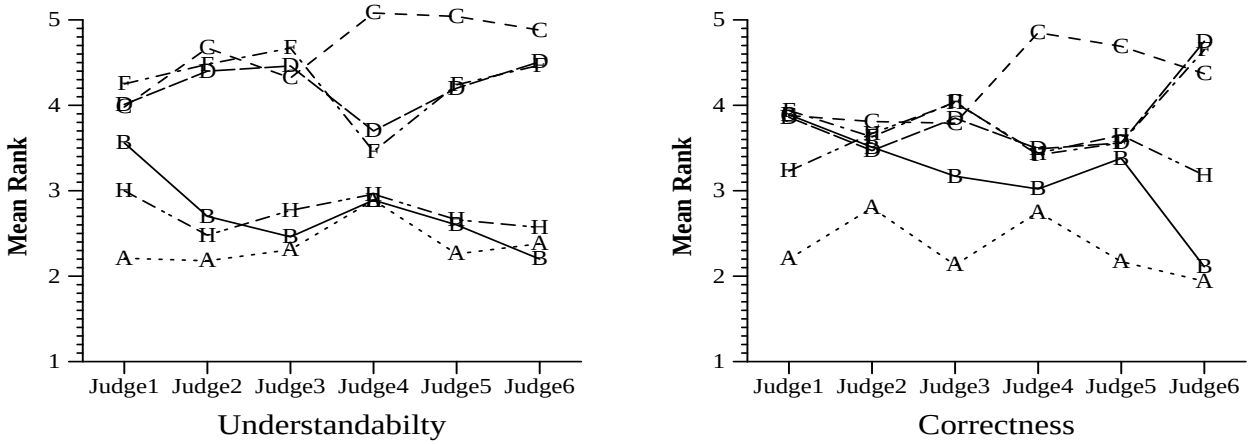

Figure 4: Relative effectiveness of runs as measured by mean rank.

the justification (as in runs D and F) was criticized as redundant. Generic phrases such as "there is a relation between" and "there is a match" were worse than useless: judges assigned no expository value to such assertions and penalized them as clutter.

Judges were also adverse to the use of system internals and jargon in the explanations. Some systems reported scores computed from WordNet (Fellbaum, 1998) or DIRT (Lin and Pantel, 2001). Such reports were penalized since the judges did not care what WordNet or DIRT are, and if they had cared, had no way to calibrate such a score. Similarly, linguistic jargon such as 'polarity' and 'adjunct' and 'hyponym' had little meaning for the judges.

Such qualitative feedback from the judges provides useful guidance to system builders on ways to explain system behavior. A broader conclusion from the justifications subtask is that it is premature for a quantitative evaluation of system-constructed explanations. The community needs a better understanding of the overall goal of justifications to develop a workable evaluation task. The relationships captured by many RTE entailment pairs are so obvious to humans (e.g., an inventor creates, a niece is a relative) that it is very unlikely end users would want explanations that include this level of detail. Having a true user task as a target would also provide needed direction as to the characteristics of those users, and thus allow judges to be more effective surrogates.

4 Conclusion

The RTE-3 extended task provided an opportunity to examine systems' abilities to detect contradiction and to provide explanations of their reasoning when making entailment decisions. True contradiction was rare in the test set, accounting for approximately 10% of the test cases, though it is not possible to say whether this is a representative fraction for the text sources from which the test was drawn or simply a chance occurrence. Systems found detecting contradiction difficult, both missing it when it was present and finding it when it was not. Levels of human (dis)agreement regarding entailment and contradiction are such that test sets for a three-way decision task need to be substantially larger than for binary decisions for the evaluation to be both reliable and sensitive.

The justification task as implemented in RTE-3 is too abstract to make an effective evaluation task. Textual entailment decisions are at such a basic level of understanding for humans that human users don't want explanations at this level of detail. User backgrounds have a profound effect on what presentation styles are acceptable in an explanation. The justification task needs to be more firmly situated in the context of a real user task so the requirements of the user task can inform the evaluation task.

Acknowledgements

The extended task of RTE-3 was supported by the Disruptive Technology Office (DTO) AQUAINT program. Thanks to fellow coordinators of the task, Chris Manning and Dan Moldovan, and to the participants for making the task possible.

References

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Lecture Notes in Computer Science*, volume 3944, pages 177–190. Springer-Verlag.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The 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. Association for Computational Linguistics.

Dekang Lin and Patrick Pantel. 2001. DIRT —Discovery of inference rules from text. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD-01)*, pages 323–328.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 200–207, July.

Ellen M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716.