*Recognizing Entailment in Intelligent Tutoring Systems**

RODNEY D. NIELSEN[1,2], WAYNE WARD[1,2] and JAMES H. MARTIN[2]

[1] *Boulder Language Technologies, 2960 Center Green Ct, Boulder, CO 80301, USA*
[2] *Department of Computer Science,* [2] *Institute of Cognitive Science*
[2] *The Center for Computational Language and Education Research*
University of Colorado, Campus Box 594, Boulder, Colorado 80309-0594, USA
Rodney.Nielsen, Wayne.Ward, James.Martin @Colorado.edu

Abstract

This paper describes a new method for recognizing whether a student's response to an automated tutor's question entails that they understand the concepts being taught. We demonstrate the need for a finer grained analysis of answers than is supported by current tutoring systems or entailment databases and describe a new representation for reference answers that addresses these issues, breaking them into detailed facets and annotating their entailment relationships to the student's answer more precisely. Human annotation at this detailed level still results in substantial interannotator agreement, 86.2%, with a Kappa statistic of 0.728. We also present our current efforts to automatically assess student answers, which involves training machine learning classifiers on features extracted from dependency parses of the reference answer and student's response and features derived from domain independent lexical statistics. Our system's performance, as high as 75.5% accuracy within domain and 68.8% out of domain, is very encouraging and confirms the approach is feasible. Another significant contribution of this work is that it represents a significant step in the direction of providing domain independent semantic assessment of answers. No prior work in the area of tutoring or educational assessment has attempted to build such domain independent systems. They have virtually all required hundreds of examples of learner answers for each new question in order to train aspects of their systems or to hand craft information extraction templates.

1 Introduction

Truly effective dialog and pedagogy in an Intelligent Tutoring System (ITS) will only be achievable when systems are able to recognize the detailed entailment relationships between a student's answer and the desired conceptual understanding. However, most state of the art ITSs simply assess student answers as a whole, classifying them as correct (the student's answer fully entails an understanding of the target concepts) or incorrect with no indication of which facets (fine-grained semantic components) of the concept the student contradicted, left unaddressed, appeared to understand, etc. Furthermore, virtually all current ITSs require a significant investment of labor to develop not just domain, but question dependent logic, parsers, knowledge structures, etc. These systems range from Finite State Machines and scripted dialogues (c.f., Pon-Barry, Clark, Schultz, Bratt and Peters 2004) at the most rigid and labor intensive end,

* This work was partially funded by NSF Awards 0551723, DRL-0733323, SBE-0518699 and by IES Award R305B070434. We would like to thank everyone who has contributed to this project with advice and support, particularly Martha Palmer and our annotators, and the anonymous reviewers, whose advice greatly improved the paper.

through Latent Semantic Analysis based systems (c.f., Graesser, Hu, Susarla, Harter, Person, Louwerse and Olde 2001), which are more flexible in handling unconstrained input, but are incapable of effectively assessing short answers or pinpointing the problems in a student's response, to hybrid machine learning based systems (c.f., Rosé, Roque, Bhembe and VanLehn 2003), which generally require 100-500 example student responses to train a new classifier *for each new question* the system is expected to handle. Furthermore, much of the parsing in the latter system is dependent on domain specific, hand coded rules, in order to capture the semantics of the domain lexicon and language.

Many other ITS researchers are striving to provide more refined learner feedback (e.g., Aleven, Popescu and Koedinger 2001; Makatchev, Jordan and VanLehn 2004). However, they too are developing very domain dependent approaches, requiring a significant investment in hand crafted logic representations, parsers, knowledge based ontologies, and dialog control mechanisms. Simply put, these domain dependent techniques will not scale to the task of developing general purpose Intelligent Tutoring Systems and will not enable the long term goal of effective unconstrained interaction with learners or the pedagogy that requires it.

There is also a small, but growing, body of research in the area of scoring constructed (free text) responses to short answer questions (e.g., Callear, Jerrams-Smith and Soh 2001; Leacock and Chodorow 2003; Mitchell, Russell, Broomhead and Aldridge 2002; Sukkarieh, Pulman and Raikes 2003). In general, short answer constructed response scoring systems are designed for large scale assessment tasks and do not provide feedback regarding the specific aspects of answers that are correct or incorrect, but merely output a raw score. Again, these approaches all require in the range of 100-500 example student answers for each *new* test question to assist in the creation of information extraction (IE) patterns or to train a classifier.

The work described in this paper represents a departure from previous ITS strategies to assess whether understanding is entailed by the student's response. First, rather than strictly checking whether the student's answer is a paraphrase of or entails the reference answer as a whole, we break the target conceptual knowledge down into fine grained facets, derived roughly from the typed dependencies in a parse of the reference answer, and check whether an understanding of these facets is entailed. This allows us to pinpoint the facet of the reference answer that the student contradicted or did not address. Second, rather than simply label the reference answer facet as being entailed or not, we provide a finer grained annotation to more precisely indicate the entailment relationship between the student's answer and that facet of the reference answer.

The paper begins with an overview of relevant prior work in paraphrasing and entailment recognition. We then highlight our current efforts to achieve more robust semantic understanding, including the development of a new evaluation framework, a large corpus to support the development, and our algorithms to detect the relationships between phrases that entail an understanding of a tutored concept and those that do not. We present our current results, an error analysis, and issues requiring further research. We also discuss the application of this new paradigm to recognition of textual entailment outside the domain of ITSs.

2 Related Prior Work on Paraphrasing and Entailment

In recent years, there has been a tremendous increase in interest in the areas of paraphrase acquisition and textual entailment recognition. Paraphrasing is the most common means for a learner to express a correct answer in an alternative form and Burger and Ferro (2005) note that even in the Pascal Recognizing Textual Entailment (RTE) challenge (Dagan, Glickman and Magnini 2005), 94% of the development corpus consisted of paraphrases, rather than what they considered true entailments. The target of much work on paraphrasing is acquiring IE patterns and fact based question answering (Agichtein and Gravano 2000; Ravichandran and Hovy 2002; Sudo, Sekine and Grishman 2001) and assumes there are several common ways of expressing the same information, (e.g., when and where someone was born). Since fact based questions are also common in testing environments, these techniques could be useful in tutoring systems. However, we are concerned more with questions that promote deeper reasoning than simple recollection and these techniques, at minimum, will require significant modifications.

Much of the remaining research on paraphrase acquisition presupposes parallel corpora (e.g., Barzilay and McKeown 2001; Pang, Knight and Marcu 2003) or comparable corpora known to cover the same news topics (e.g., Barzilay and Lee 2003; Dolan, Quirk and Brockett 2004). The parallel corpora are aligned at the sentence level, with sentence pairs considered to be paraphrases, and often lattices are created to form a database of potential paraphrasing transformations. Following these techniques, Barzilay and McKeown found that only around 35% of the lexical paraphrases they extracted were synonyms; the remaining were 32% hypernyms, 18% siblings of a hypernym, 5% from other WordNet relations, and 10% were not related in WordNet, emphasizing the need for relatively loose entailment recognition. Other techniques also exist for lexical paraphrase acquisition, which do not rely on parallel corpora (e.g., Glickman and Dagan 2003).

Research in the area of entailment also has much to offer. Lin and Pantel (2001) extract inference rules from text by looking for dependency parse patterns that share common argument fillers; thus, taking Harris' Distributional Hypothesis that words occurring in similar contexts tend to have similar meanings and extending it to dependency paths or phrases. The difficulty in directly applying this work to the task of answer assessment is that the inference rules extracted do not tend to have broad coverage. This is evidenced by the fact that several researchers participating in the Pascal RTE challenge made use of Lin and Pantel's patterns and did not find them to improve performance (e.g., Braz, Girju, Punyakanok, Roth and Sammons 2005; Raina, Haghighi, Cox, Finkel, Michels, Toutanova, MacCartney, de Marneffe, Manning and Ng 2005).

The RTE challenge has brought the issue of textual entailment before a broad community of researchers in a task independent fashion. The challenge requires systems to make yes-no (or unknown in a RTE3 pilot task) judgments as to whether a human reading a text *t* would typically consider a second, hypothesis, text *h* to most likely be true. The following example shows a typical *t-h* pair from the RTE challenge. In this example, the entailment decision is *no* – and that is similarly the extent to which training data is annotated. There is no indication of whether some facets of *h*

are addressed in *t* (as they are in this case) or conversely, which facets are not discussed or are explicitly contradicted.

t: At an international disaster conference in Kobe, Japan, the U.N. humanitarian chief said the United Nations should take the lead in creating a tsunami early-warning system in the Indian Ocean.

h: Nations affected by the Asian tsunami disaster have agreed the UN should begin work on an early warning system in the Indian Ocean.

Submitters to the RTE challenge take a variety of approaches including lexical similarity approaches (e.g., Glickman, Dagan and Koppel 2005), lexical-syntactic feature similarity (e.g., Nielsen, Ward and Martin 2006), syntactic/description logic subsumption (e.g., Braz et al. 2005), graph matching with semantic roles (Raina et al. 2005), logical inference (e.g., Tatu and Moldovan 2007), discourse commitment – assertion, presupposition, and conversational implicature – strategies (Hickl and Bensley 2007), and numerous others. In the following sections, we build on many of these techniques and, more importantly to our cause, we develop a more expressive paradigm and representation framework for recognizing detailed entailment relationships at a fine grained level. This framework is applicable to a wide variety of application areas and, in automated tutoring systems, we believe it will lead to more effective dialog and improved learning.

3 Annotating a Corpus with Fine Grained Entailments

3.1 The Representation

The goal of the representation described here is to facilitate domain independent assessment of student responses to questions in the context of a known reference answer and to perform this assessment at a level of detail that will enable more effective ITS dialog. We have two key criteria for this representation: 1) it must be at a level that facilitates detailed assessment of the learner's understanding, indicating exactly where and in what manner the answer did not meet expectations and 2) the representation and assessment should be learnable by an automated system.

Rather than have a single yes or no entailment decision for the reference answer as a whole, we instead break the reference answer down into what we consider to be its finest grained compositional facets. This roughly translates to the set of triples composed of labeled dependencies in a dependency parse of the reference answer, but we use the word *facet* throughout this paper to generically refer to some part of a text's (or utterance's) meaning. To facilitate system testing and the development of a meaningful entailment corpus, this decomposition was performed manually in this work, but the process will be automated in the future. The following illustrates how a simple reference answer (1) is decomposed into the answer facets (1a-d) derived from its dependency parse. The dependency parse tree is shown in Figure 1. As can be seen in 1b and 1c, the dependencies are augmented by thematic roles (c.f., Kipper, Dang and Palmer 2000) (e.g., Agent, Theme, Cause, Instrument, etc).

(1) A long string produces a low pitch.
(1a) NMod(string, long)
(1b) Agent(produces, string)
(1c) Product(produces, pitch)

(1d) NMod(pitch, low)

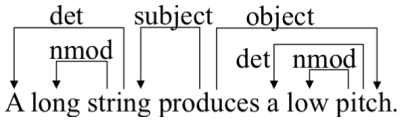

Fig. 1. Dependency parse tree for example reference answer

Breaking the reference answer down into fine grained facets provides the tutor's dialog manager with a much finer grained assessment of the student's response, but a simple yes or no entailment at the facet level still lacks semantic expressiveness with regard to the relation between the student's answer and the facet in question. For example, did the student contradict the facet or completely fail to address it? We also need to break the annotation labels into finer levels in order to specify more clearly these relationships. These two issues – breaking the reference answer into fine grained facets and utilizing more expressive annotation labels – are the emphasis of the present work; the associated representation is central to the corpus and our entailment system design.

3.2 The Corpus

Our work focuses on the critical years of learning to comprehend text, K-6. We are currently in the process of implementing our ITS, and lacking live tutoring transcripts, we acquired data gathered from 3rd-6th grade students utilizing the Full Option Science System (FOSS), upon which our ITS is based. FOSS is a proven system that has been in use across the United States for over a decade (Lawrence Hall of Science 2005). Assessment is a major FOSS research focus, of which the Assessing Science Knowledge (ASK) project is a key component (Lawrence Hall of Science 2006). FOSS includes sixteen diverse science modules, spanning physical science, earth science, life science, space science, scientific reasoning and technology, and for each module, ASK includes a set of assessment questions with reference answers.

We used 287 constructed response questions taken from the ASK assessments covering their sixteen science teaching and learning modules. These questions had expected responses ranging in length from moderately short verb phrases to several sentences and could be assessed objectively, (representative questions with their reference answers and an example student answer are shown in Table 1). We generated a corpus from a random sample of the students' handwritten responses to these questions. The only special transcription instructions were to fix spelling errors (since these would be irrelevant in our target spoken dialog environment), but not grammatical errors (which would still be relevant), and to skip blank answers and non-answers similar in nature to *I don't know*. We transcribed approximately 15,400 student responses (roughly 100 per question for three test set modules (Environment, Human Body and Water) and forty per question for the remaining thirteen modules – section 4.3 details the test sets).

Table 1 Sample ASK Qs with reference (R) and example student (A) answers.

Human Body	Q:	Dancers need to be able to point their feet. The tibialis is the major muscle on the front of the leg and the gastrocnemius is the major muscle on the back of the leg. Describe how the muscles in the front and back of the leg work together to make the dancer's foot point.
	R:	The muscle in the back of the leg (the gastrocnemius) contracts and the muscle in the front of the leg (the tibialis) relaxes to make the foot point.
	A:	The back muscle and the front muscle stretch to help each other pull up the foot.
Structure of Life	Q:	Why is it important to have more than one shelter in a crayfish habitat with several crayfish?
	R:	Crayfish are territorial and will protect their territory. The shelters give them places to hide from other crayfish. [Crayfish prefer the dark and the shelters provide darkness.]
	A:	So all the crayfish have room to hide and so they do not fight over them.
Magnetism and Electricity	Q:	Lee has an object he wants to test to see if it is an insulator or a conductor. He is going to use the circuit you see in the picture. Explain how he can use the circuit to test the object.
	R:	He should put one of the loose wires on one part of the object and the other loose wire on another part of the object (and see if it completes the circuit).
	A:	You can touch one wire on one end and the other on the other side to see if it will run or not.

3.3 The Annotation

Manual corpus annotation is a two step process.[1] First, each reference answer (as specified by the ASK research team) is decomposed by hand into its constituent facets. Then for each student answer, the associated reference answer facets are annotated to describe whether and how they are addressed by the student.

3.3.1 Reference Answer Facet Extraction

The reference answer facets are roughly extracted from the relations in a syntactic dependency parse (Nivre, Hall, Nilsson, Chanev, Eryigit, Kubler, Marinov and Marsi 2007) and a shallow semantic parse (Gildea and Jurafsky 2002). These relations are then modified to increase the semantic value of the facets, as described in the following paragraphs. Figure 2 shows a typical dependency parse and the revised parse for a reference answer fragment that illustrates several of these modifications. Example 2 shows the decomposition of this same answer fragment into its constituent facets along with their glosses. The rationale for these modifications is that they will facilitate more reasonable comparisons between the student and reference answers as elaborated below.

(2) The brass ring would not stick to the nail because the ring is not iron.
(2a) NMod(ring, brass)
(2a') The ring is brass.
(2b) Theme_not(stick, ring)
(2b') The ring does not stick.

[1] The corpus can be freely downloaded via the *Resources* page linked to the first author's homepage at http://www.bltek.com/main-link/

(2c) Destination_to_not(stick, nail)
(2c') Something does not stick to the nail.
(2d) Be_not(ring, iron)
(2d') The ring is not iron.
(2e) Cause_because(2b-c, 2d)
(2e') 2b and 2c are caused by 2d.

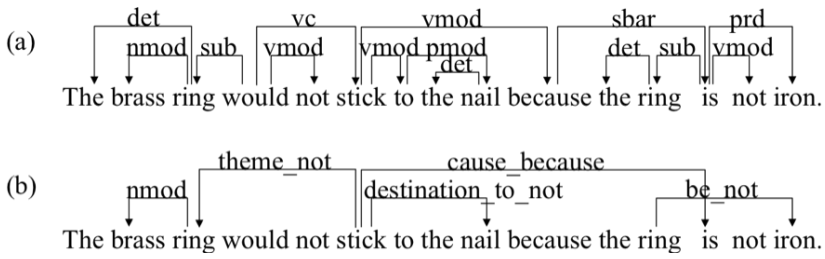

Fig. 2. Example dependency parse tree (a) and final representation (b)

First, auxiliary verbs and their modifiers are reattached to the associated main verbs (e.g., in Figure 2, *would*, *ring* and *not* are reattached to *stick*). Then relations that carry little semantic content are removed (e.g., dependencies involving most determiners and auxiliaries). The rationale for these changes is that a statistical analysis to determine whether the learner answer entails the dependency VC(*would, stick*) is unlikely to yield more informative results then simply assessing the strictly lexical entailment of *stick*, with similar results for SUB(*would, ring*) versus just ring; whereas, SUB(*stick, ring*) does add semantic value over its lexical constituents, requiring that the student express the full concept involving their relation.

The parses were also modified to incorporate prepositions, copulas, terms of negation, and similar terms into the dependency type labels. In Figure 2, these changes included the incorporation of *to*, *because*, *be*, and the two instances of *not* into their associated dependencies. The rationale for these changes is similar to that given above, (e.g., assessing the entailment of VMod(*stick, to*) provides little value over the lexical entailment of *stick*; whereas, Destination_to_not(*stick, nail*) carries significantly more semantics and will have a greater effect on entailment decisions). Finally, we replace the standard dependency type labels with thematic roles where they are relevant. This is predominantly verb arguments (the *theme*, *destination* and *cause* in Figure 2), but occasionally includes noun modifiers (e.g., in *The water on the floor had a much larger surface area*, we extract the facet Location_on(*water, floor*)). We also add semantic role relations that are not derivable from a typical dependency parse. For example, in the sentence '*As it freezes the water will expand and crack the glass*', the relation between *water* and *crack* is not included in the output of most dependency parsers, but we do extract a facet representing the Agent role that *water* plays in this event.

The modifications described here result in much more compact and meaningful facets. In Figure 2, the number of dependencies is reduced from fifteen (with the root) to five, including the replacement of the four semantically impoverished dependencies Det(*ring, the*), Sub(*is, ring*), VMod(*is, not*), and Prd(*is, iron*) with the single more expressive facet Be_not(*ring, iron*). Example 3 illustrates the decomposition of another reference answer into its constituent parts along with their glosses.

(3) The string is tighter, so the pitch is higher.
(3a) Be(string, tighter)
(3a') The string is tighter.
(3b) Be(pitch, higher)
(3b') The pitch is higher.
(3c) Cause_so(3b, 3a)
(3c') 3b is caused by 3a

3.3.2 Reference Answer Facet Entailment Annotation

After analyzing much of the Physics of Sound data, we settled on the eight annotation labels noted in Table 2 (Nielsen and Ward 2007). Descriptions of where each annotation label applies and some of the most common annotation issues were detailed with several examples in the annotator guidelines and are summarized below. We only annotate a student's answer relative to the constituent facets of the associated reference answer, (i.e., for a given student answer, we annotate each facet of the corresponding reference answer with a label indicating the student's apparent understanding of that facet; if the reference answer has six total facets, every corresponding student answer will be annotated with six labels, one per facet). If the student also discusses concepts not addressed in the reference answer, those points are not annotated regardless of their quality or accuracy.

Table 2. Facet Annotation Labels

Label	Brief Definition
Assumed	Reference answer facets that are assumed to be understood a priori based on the question
Expressed	Any Reference answer facet directly expressed or inferred by simple reasoning
Inferred	Reference answer facets inferred by pragmatics or nontrivial logical reasoning
Contra-Expr	Reference answer facets directly contradicted by negation, antonymous expressions and their paraphrases
Contra-Infr	Reference answer facets contradicted by pragmatics or complex reasoning
Self-Contra	Reference answer facets that are both contradicted and implied (self contradictions)
Diff-Arg	Reference answer facets where the core relation is expressed, but it has a different modifier or argument
Unaddressed	Reference answer facets that are not addressed at all by the student's answer

Example 4 shows a question and a fragment of its reference answer broken down into its constituent facets with an indication of whether the facet is assumed to be understood a priori. A corresponding student answer is shown in (5) along with its final annotation in 4a'-c'. It is assumed that the student understands that the pitch is higher a priori (reference answer facet 4b), since this is given in the question (… Write a note to David to tell him why the pitch gets higher rather than lower) and similarly it is assumed that the student will be explaining what has the causal effect of producing this higher pitch (facet 4c). Therefore, unless the student explicitly addresses these facets they are labeled Assumed.

(4) Question: After playing the FOSS-ulele, David wrote his results in his lab notebook:
I'm confused. When I pull down and tighten the string on the FOSS-ulele, then pluck the string, the pitch sounds HIGHER than it did before. But aren't I making the string longer when I pull the string? I thought a longer length produced a LOWER pitch. What's going on here?
What is causing the pitch to be higher? Write a note to David to tell him why the pitch gets higher rather than lower.
Reference Answer: The string is tighter, so the pitch is higher.

(4a) Be(string, tighter), ---

(4b) Be(pitch, higher), Assumed

(4c) Cause(4b, 4a), Assumed

(5) David this is why because you don't listen to your teacher. If the string is long, the pitch will be high.

(4a') Be(string, tighter), Diff-Arg

(4b') Be(pitch, higher), Expressed

(4c') Cause(4b', 4a'), Expressed

Since the student does not contradict the fact that the string is tighter (it can be both longer and tighter), we do not label this reference answer facet (4a') Contradicted. If the student's response did not mention anything about either the string or tightness, we would annotate 4a' as Unaddressed. However, the student did discuss a property of the string, *the string is long*. This parallels the reference answer facet Be(string, tighter) with the exception of a different argument to the Be relation, resulting in the annotation Diff-Arg. This indicates to the tutor that the student expressed a related concept, but one which neither implies that they understand the facet nor that they explicitly hold a contradictory belief. Often, this indicates the student has a misconception. For example, when asked about an effect on pitch, many students say things like *the pitch gets louder*, rather than higher or lower, which implies a misconception involving their understanding of pitch and volume. In this case, the Diff-Arg label can help focus the tutor on correcting this misconception. Facet 4c', expressing the causal relation between 4a' and 4b', is labeled Expressed, since the student did express a causal relation between the concepts aligned with 4a' and 4b'. The tutor then knows that the student was on track in regard to attempting to express the desired causal relation and the tutor need only deal with the fact that the cause given was incorrect.

The Self-Contra annotation is used in cases like the response in example 6, where the student simultaneously expresses the contradictory notions that the string is tighter and that there is less tension.

(6) The string is tighter, so there is less tension so the pitch gets higher.

(4a") Be(string, tighter), Self-Contra

(4b") Be(pitch, higher), Expressed

(4c") Cause(4b", 4a"), Expressed

Example 7 illustrates a case where a student's answer is labeled Inferred, (facet 7b). In this case, the decision requires pragmatic inferences, applying the Gricean maxims of Relation, be relevant – why would the student mention vibrations if they did not know they were a form of movement – and Quantity, do not make your contribution more informative than is required (Grice, 1975).

(7) Question: Kate said: "An object has to move to produce sound." Do you agree with her? Why or why not?
Reference Answer: "Agree. Vibrations are movements and vibrations produce sound."
Student Answer: Yes because it has to vibrate to make sounds.
(7a) Root(root, agree), Expressed
(7b) Be(vibration, movement), Inferred
(7c) Agent(produce, vibrations), Expressed
(7d) Product(produce, sound), Expressed

There is no compelling reason from the perspective of the automated tutoring system to differentiate between Expressed, Inferred and Assumed facets, since in each case the tutor can assume that the student understands the concepts involved. However, from the systems development perspective there are three primary reasons for differentiating between these facets and similarly between facets that are contradicted by inference versus by more explicit expressions. First, including the more difficult, often pragmatic, inferences in the training data could have negative effects on some machine learning algorithms. Having separate labels allows one to remove the more difficult inferences from the training data, thus eliminating this problem. The second rationale is that systems hoping to handle both types of inference might more easily learn the characteristics of each class if they are distinguished. The third reason for separate labels is that it facilitates system evaluation, including the comparison of various techniques and the effect of individual features.

3.3.3 Reference Answer Facet Entailment Annotation Results

We evaluate our annotation results under three label groupings: 1) All-Labels, where all of the labels are kept separate, 2) Tutor-Labels, (the five labels that will be used by the system), where Expressed, Inferred and Assumed are combined into a single Understood category (i.e., the student understands the facet) and Contra-Expr and Contra-Infr are combined to form Contradicted, but the other labels are left as is, and 3) Yes-No, which is a two way division, Understood versus all other labels. We evaluated interannotator agreement on all double annotated data in the sixteen science modules, totaling 142,451 facet annotations. Agreement on the Tutor-Labels is 86.2%, with a Kappa statistic of 0.728 corresponding with substantial agreement (Cohen 1960). Agreement is 78.4% on All-Labels and 88.0% on the binary Yes-No decision. These also have Kappa statistics in the range of substantial agreement.

The distribution of labels is shown in Table 3. An analysis of the interannotator confusion matrix indicates that the most probable disagreement is between Inferred and Unaddressed, comprising 39% of the disagreements. The next most likely disagreements are between Expressed and the other Understood labels (Inferred and Assumed), representing 35% of the disagreements. Confusion between Expressed and Unaddressed is also considerable, representing 10% of the annotator disagreements.

Table 3. Distribution of annotation labels (145,911 facet annotations)

Label	Count	% of Total	Count	% of Total
Assumed	37,076	25.4		
Expressed	31,555	21.6	89,105	61.1
Inferred	20,474	14.0		
Contra-Expr	1,426	1.0	2,482	1.7
Contra-Infr	1,056	0.7		
Self-Contra	86	0.1		0.1
Diff-Arg	1,780	1.2		1.2
Unaddressed	52,458	36.0		36.0

4 Robust Entailment Recognition in Intelligent Tutoring

In this section, we describe our initial system for entailing or assessing a student's understanding of individual reference answer facets; we present results, an error analysis and related plans for future work; and we discuss an adaptation to the RTE challenge task and its results. A high level description of the overall procedure is as follows. We start with hand generated reference answer facets, similar to typed dependency triples (see section 3.3.1). We generate automatic parses for the reference answers and the student answers and automatically modify these parses per our desired representation summarized in section 3.1 (also see section 3.3.1 for further details). Then for each manually generated reference answer facet, we extract features indicative of the student's understanding of that facet. Finally, we train a machine learning classifier on the training data and use it to classify the unseen examples in the test set, assigning one of the five Tutor-Labels separately for each reference answer facet to indicate the student's understanding of that facet.

4.1 Preprocessing and Representation

Many of the machine learning features utilized here are based on document cooccurrence counts. Rather than use the web as our corpus (as did Turney (2001) and Glickman et al. (2005), who generate analogous similarity statistics), we use three publicly available corpora (English Gigaword, The Reuters corpus, and Tipster) totaling 7.4M articles and 2.6B indexed terms. These corpora were utilized because they were readily available and already indexed. However, they are all drawn predominantly from the news domain, making them less than ideal for assessing students' answers to science questions. This will be addressed in the future by indexing more relevant information drawn from the web. These corpora were indexed and searched with Lucene, utilizing their PorterStemFilter to replace the surface form of words with their lexical stem (the index excluded only three words, {*a, an, the*}).

Before extracting features, we generate dependency parses of the reference answers and student answers using MaltParser (Nivre et al. 2007). These parses are automatically modified per our desired representation by reattaching auxiliary verbs and their modifiers to the associated main verbs and incorporating prepositions, copulas, terms of negation and similar terms into the dependency relation labels (see Section 3.3.1 and Figure 2; however, we have not trained or utilized a thematic role labeler yet). As

described in Section 3.3.1, these modifications increase the likelihood that terms carrying significant semantic content are joined by dependencies that will be the focus of feature extraction.

4.2 Machine Learning Features

We investigated a variety of linguistic features and settled on the features summarized in Table 4, informed by training set cross validation results. Many of the features dropped provided significant value over the simple lexical baseline, but did not improve on the more informative features described here. The features assess lexical similarity via lexical entailment probabilities following (Glickman et al. 2005) and lexical stem matches. They include syntactic information such as part of speech, relevant dependency or facet relation types and dependency path edit distances. Other features include polarity among other things (see Table 4 for details).

Table 4. Machine learning feature descriptions

LEXICAL FEATURES
GOV/MOD_MLE: The lexical entailment probabilities for the reference answer facet governor (Gov) and modifier (Mod) following (Glickman et al. 2005)
GOV/MOD_MATCH: True if the Gov (Mod) stem has an exact match in the learner answer
SUBORDINATE_MLES: The lexical entailment probabilities for the primary constituent facets' governors and modifiers when the reference facet represents a relation between higher level propositions.
SYNTACTIC FEATURES
GOV/MOD_POS: The part of speech tags for the reference facet's governor and modifier
FACET/ALIGNEDDEP_RELTN: The type labels of the reference facet and aligned learner answer dependency (dependency alignments are based on cooccurrence MLEs as with words, i.e., they estimate the likelihood of seeing the reference answer dependency in a document given it contains the learner answer dependency (Nielsen et al. 2006))
DEP_PATH_EDIT_DIST: The edit distance between the dependency path connecting the reference facet's Gov and Mod (not necessarily a single step due to parser errors) and the path connecting the aligned terms in the learner answer. Paths include the dependency relations generated in our modified parse with their attached prepositions, negations, etc, the direction of each dependency, and the POS tags of the terms on the path. The calculation applies heuristics to judge the similarity of each part of the path (e.g., dropping a subject has a much higher cost than dropping an adjective). Alignment for this feature is based on which set of terms in an N-best list (N=5 in the present experiments) for the Gov and Mod result in the smallest edit distance. The N-best list is generated based on the lexical entailment probabilities above.
OTHER FEATURES
CONSISTENT_NEGATION: True if the aligned learner dependency path has the same number of negations as the reference answer facet.
RA_CW_CNT: The number of content words in the reference answer, motivated by the fact that longer answers are more likely to result in spurious alignments.

4.3 Classification Approach

The feature data was split into a training set and three test sets. The first test set, *Unseen Modules*, consists of all the data from three of the sixteen science modules (Environment, Human Body and Water), providing what is loosely a domain independent test set of topics not seen in the training data. The second test set, *Unseen Questions*, consists of all the student answers associated with twenty two randomly selected questions from the 233 questions in the remaining thirteen modules, providing a question

independent test set from the same topics seen in the training data. The third test set, *Unseen Answers*, was created by withholding four random learner answers from each of the remaining questions, with the remainder comprising the training set.

Today's ITSs know in advance the questions, reference answers, and what information a student should be *assumed* to understand a priori (e.g., based on the question context). This is equivalent to the Unseen Answers test set; so, in that scenario, it makes less sense to include the facets that are labeled Assumed in the test set. The long term goal is for our ITS to generate its own questions, in which case it would be useful to classify facets that should be assumed understood a priori. However, this requires special logic and feature sets that we have not yet implemented, so all Assumed facets were withheld from the data sets. This selection resulted in a total of 54,967 training examples, 30,514 examples in the Unseen Modules test set, 6,699 in the Unseen Questions test set and 3,159 examples in the Unseen Answers test set.

We evaluated several machine learning algorithms and C4.5 (Quinlan 1993) achieved the best results in cross validation on the training data. Therefore, we used it to obtain results for this new task of automatically annotating fine grained reference answer facets with detailed entailment classifications.

4.4 Results

Table 5 shows the classifier's accuracy in 10-fold cross validation on the training set as well as on each of our test sets. The columns first show two simpler baselines, the accuracy of a classifier that always chooses the most frequent label in the *training* set (Unaddressed), and the accuracy based on a lexical decision that chooses Understood if both the reference answer facet's governing term and its modifier are present in the learner's answer and outputs Unaddressed otherwise, (we also tried thresholding the product of their lexical entailment probabilities following Glickman et al. (2005), who achieved the best results in the first RTE challenge, but this gave virtually the same results as the word matching baseline for this dataset). The column labeled *Table 4 Features* presents the results of our classifier and *Reduced Training* is described in the Discussion section which follows. These results are for predicting the five possible *Tutor Labels* that will drive the dialogue provided by the tutor: Understood, Contradicted, Self-Contra, Diff-Arg, and Unaddressed.

Table 5. Tutor-Labels classifier accuracy on non-Assumed reference answer facets

	Majority Label	Lexical Baseline	Table 4 Features	Reduced Training
Training Set 10x CV	54.6	59.7	**77.1**	
Unseen Answers	51.1	56.1	**75.5**	
Unseen Questions	58.4	63.4	61.7	**66.5**
Unseen Modules	53.4	62.9	61.4	**68.8**

4.5 Discussion

This is a novel paradigm and a new dataset, and these results are very promising. The accuracy on the Tutor-Labels is 24.4%, 3.3%, and 8.0% over the most frequent class baseline for Unseen Answers, Questions, and Modules respectively. Accuracy on Un-

seen Answers is 19.4% better than the lexical baseline. However, this simpler baseline outperformed our classifier on the other two test sets. Results on the Yes-No Labels (not shown), predicting that the tutor should provide some form of remediation, follow a similar trend.

It seemed probable that the decision tree may have over fit the data due to bias in the data itself; specifically, since many of the students' answers are very similar when considering just the text that is relevant to a given reference answer facet, there are likely to be large clusters of nearly identical feature-class pairings, which could result in classifier decisions that do not generalize as well to other questions or domains. This bias is not problematic when the test data is very similar to the training data, as is the case for our Unseen Answers test set, but would negatively affect performance on less similar data, such as our Unseen Questions and Modules. To test this hypothesis, we reduced the number of examples in our training set to about 8,000, which would result in fewer of these dense clusters, and retrained the classifier. The result for Unseen Questions, shown in the *Reduced Training* column, was an improvement of 4.8%. Given this promising improvement, we attempted to find the optimal training set size through cross-validation on the training data. Specifically, we iterated over the training science modules holding one module out, training on the other 12 and testing on the held out module. We analyzed the learning curve varying the number of randomly selected examples per facet and found the optimal accuracy for training set cross-validation by averaging the results over all the modules. We then trained a classifier on that number of random examples per facet in the training set and tested on the Unseen Modules test set. The result was an increase in accuracy of 7.4% over training on the full training set. In future work, we will investigate other more principled techniques to avoid this type of over-fitting, which we believe is somewhat atypical, and analyze its cause in more detail.

Results on the Unseen Modules test set are suggestive of performance that might be achievable in other domains, such as general reading comprehension. In this case, the system was trained on data from areas such as Physics of Sound and Electricity and Magnetism, and then tested on very different topics such as the Human Body and Water. While this work includes extensive hand annotation, no *training* data needed to be annotated in these test domains and the system was not modified in any way to handle the new domains (annotation of the three associated modules was strictly for testing). Furthermore, feature vectors were extracted from term and dependency similarities found in news wire corpora, rather than domain specific science corpora. This work is not completely domain independent, since the reference answer facets were extracted by hand. However, this process was based closely on techniques that have been proven to be learnable by systems (dependency parsing and shallow semantic parsing). Indeed many of our features are derived from the automatic generation of a representation very similar to that of the reference answer facets. Still, it is possible that, if the style of questions and reference answers varies from that of our corpus, it could result in a drop in the domain independent performance. Future work includes the automatic generation of the reference answer facets, at which point no hand annotation would be required to handle new questions or domains. The adaptation of this system to the RTE challenge task (see section 4.7) represents a significant step in this

direction and provides evidence that the system should perform approximately as well when the tutored subject is very different.

It is worth emphasizing that the annotation (system and human) is strictly relative to the reference answer facets (i.e., for each student answer, each facet of the reference answer is annotated to indicate the perceived student understanding of it). Student answer facets are not annotated separately to indicate their relevance and, thus, the system is not being unduly credited for identifying irrelevant student verbiage. Future work includes the detection of misconceptions that are independent of the reference facets; however, it is rare for students to express misconceptions that do not result in at least one reference answer facet being labeled Contradicted or Diff-Arg.

4.6 Error Analysis

In order to focus future work on the areas most likely to benefit the system, an error analysis was performed based on the results of 13-fold cross-validation on the training data. For each module, we trained on twelve science modules and tested on the data in the remaining, held-out module, effectively simulating the Unseen Modules test condition. We only considered the subsets of errors that were most likely to lead to short-term system improvements. This included only Expressed and Unaddressed facets where the annotators agreed unanimously. Contradictions were excluded since there was almost no attempt to handle these in the present system. This must be addressed in future work, since despite their somewhat infrequent occurrence, it is critical that the tutor recognize and address these contradictory beliefs as early as possible. We discuss Expressed facets in the next section of the paper and Unaddressed in the following section.

4.6.1 Errors in Expressed Facets

Without examining each example relative to the decision tree that classified it, it is not possible to know exactly what caused the errors. The analysis here simply indicates what factors are involved in inferring whether the reference answer facets were understood and what relationships exist between the student answer and the reference answer facet. We analyzed 100 random examples of errors where annotators considered the facet Expressed and the system labeled it Unaddressed. We group the potential error factors seen in the data, listed in order of frequency, according to issues associated with paraphrases, logical inference, pragmatics, and preprocessing errors.

Paraphrase issues, broadly construed, are subdivided into four main categories: coreference resolution, lexical substitution, syntactic alternation and phrase-based paraphrases. Our results in this area are in line with (Bar-Haim, Szpektor and Glickman 2005), who considered which inference factors are involved in proving textual entailment in the RTE challenge. Three coreference resolution factors combined are involved in nearly 30% of the errors. Students use on average 1.1 pronouns per answer, generally referring to key entities. Fifteen of the errors involved a pronoun, including eleven uses of *it*. In twelve errors, the student utilized a coreferring common noun or adjective from the question or reference answer.

As a group, the simple lexical substitution categories (synonymy, hypernymy, hyponymy, meronymy, derivational changes, and other lexical paraphrases) appear more often in errors than any of the other factors with around thirtyfive occurrences.

Roughly half of these relationships should be detectable using broad coverage lexical resources. However, many of these lexical paraphrases are not necessarily associated in lexical resources such as WordNet or VerbNet (e.g., neither resource includes a connection between *have* and *contain*). Concept definitions account for an additional fourteen issues that might be addressed by lexical resources.

Vanderwende, Coughlin and Dolan (2005) found that 34% of the Recognizing Textual Entailment Challenge test data could be handled by recognizing simple syntactic variations. However, while syntactic variation is certainly common in the kids' data, it did not appear to be the primary factor in any of the system errors. Most of the remaining paraphrase errors were classified as involving phrase-based paraphrases (e.g., *in the middle* versus *halfway between*). Six related errors essentially involved negation of an antonym, (e.g., *no one has the same fingerprint* versus *everyone has a different print*). Paraphrase recognition is an area that we intend to invest significant time in future research.

The next most common issues after paraphrases were deep or logical reasoning (e.g., recognizing that two sounds must be *very different* in the case that *…it is easy to discriminate* [between them]…) and then pragmatics (e.g., recognizing that saying *Because the vibrations* implies that a rubber band is vibrating given the question context). These two factors were involved in nearly 40% of the errors. Finally, the remaining errors were largely the result of preprocessing issues (e.g., data normalization: *3* versus *three*).

While we believe improving the parser output will result in higher accuracy by the entailment classifier, there was little evidence to support this in the parses examined in this error analysis. We only checked parses when the dependency path features looked wrong and it was somewhat surprising that the classifier made an error and only two of these classification errors were associated with parser errors. However, better parses should lead to more reliable (less noisy) features, which in turn will allow the machine learning algorithm to more easily recognize which features are the most predictive.

It should be emphasized that over half of the errors in Expressed facets involved more than one of the fine-grained factors discussed here. For example, to recognize the child understands a tree is blocking the sunlight based on the answer *There is a shadow there because the sun is behind it and light cannot go through solid objects...*, requires resolving the *solid object* mentioned to the *tree*, and then recognizing that *light cannot go through* [*the tree*] entails the tree blocks the light.

4.6.2 Errors in Unaddressed Facets

Unlike the errors in Expressed facets, a number of the examples here appeared to be questionable annotations. For example, given the student answer *Because the darker the color the faster it will heat up*, the annotators did not infer that the student believed the sheeting chosen was the *darkest color*.

One of the biggest sources of errors in Unaddressed facets is the result of ignoring the context of words. For example, the student answer *You could wrap the insulated wire to the iron nail and attach the battery and switch* leads to the classification of Understood for a facet indicating to *touch the nail* to a permanent magnet to turn it

into a temporary magnet, but *wrapping* the wire *to the nail* should have been aligned to a different method of making a temporary magnet.

Many of the errors in Unaddressed facets appear to be the result of antonyms having very similar statistical co-occurrence patterns (e.g., *absorbs energy* versus *reflects energy*). However, this also may be an annotation error that should have been labeled Contra-Expr rather than Unaddressed.

The biggest source of error is simply classifying a number of facets as Understood if there is partial lexical similarity and perhaps syntactic similarity as in the case of accepting *the balls are different* in place of *different girls*. However, there are also a few cases where it is unclear why the decision was made, as in an example where the system apparently trusted that the student understood a complicated electrical circuit based on the student answer *we learned it in class*.

The processes and the more informative features implied in the preceding section describing errors in Expressed facets should allow the learning algorithm to focus on less noisy features and avoid many of the errors described in this section. However, additional features will need to be added to ensure appropriate lexical and phrasal alignment, which should also provide a significant benefit here. Future plans include training an alignment classifier separate from the assessment classifier.

4.7 Application to RTE outside of Intelligent Tutoring Systems

We believe this work is applicable to recognizing textual entailment in general, beyond just the ITS domain. For example in question answering, answer facets that are entailed by multiple passages in a text can be asserted with more confidence or answers could be built up from separately entailed facets. In the remainder of this section, we describe one possible adaptation of our system to the RTE challenge to situate it within the broader RTE field.

For a number of reasons it is not straightforward to apply our system to the RTE challenge datasets nor to compare its performance with systems in the challenges. First, we are classifying reference answer *facet* entailment, not entailment of an *overall passage* (called the hypothesis, *h*, in the RTE challenge). An implementation that simply classifies *h* as entailed if each of its facets are entailed and not entailed otherwise, would perform poorly. Second, we currently start with manually generated reference answer facets and do not have such facets for *h*; furthermore, creating them would not lead to a fair comparison with other systems. To work around this issue, we instead automatically generated facets as described in section 4.1. The key difference, beyond the fact that the ASK reference answer facets are gold standard, is that the automatically generated *h* facets use dependency type labels rather than thematic role labels.

We trained a binary facet-based classifier on the ASK corpus, withholding Inferred, Self-Contradicted and Assumed facets, as these were largely beyond the scope of the RTE challenge datasets or our current system implementation. The binary decision classified facets as Entailed (Expressed) versus Not Entailed (Contradicted, Unaddressed or Diff-Arg). We then classified each system-generated facet of the RTE datasets and produced associated facet entailment probability estimates. Our feature set for this facet classifier was as described in Table 4 with two exceptions. First, we excluded the FACET_RELTN and ALIGNED_DEP_RELTN, since the RTE data was labeled

with dependency types rather than thematic role labels. Second, we excluded SUBORDINATE_MLEs, since these features were used only for the facets connecting higher-level propositions and we wanted a consistent feature set when combining results across all facets. The accuracy of this binary classifier on our Unseen Questions and Unseen Modules datasets was 80.8% and 79.2%, respectively.

We estimated accuracy on the RTE facets by assuming that all facets in entailed RTE pairs were entailed and labeling a random sample of 200 facets from the non-entailed pairs (100 classified entailed and 100 classified not-entailed). The accuracy on facets from entailed pairs was 79.8% and the estimated accuracy on facets from non-entailed pairs was 77.9%, resulting in an overall estimated RTE facet accuracy of 78.9%. This is only 0.3% below the accuracy on the ASK Unseen Modules dataset, further supporting the belief that the system should transfer to other ITS domains with relatively little degradation in performance and without the need for additional hand annotation or domain specific development. It is worth emphasizing that there was absolutely no manual generation of RTE facets, nor was there any RTE facet labeling for the purposes of training; the system was trained strictly on the children's data in the ASK corpus.

Next, we generated feature vectors for each RTE example by combining the features and entailment probability estimates over each of the facet-based classifications. Specifically, the features for the final classifier were the average, geometric mean, and lowest values over all facets for the ASK-trained classifier's facet entailment probabilities, the facet-based GOV/MOD_MLE features, and the DEP_PATH_EDIT_DIST features; the proportion of exact stem matches (from GOV/MOD_MATCH); the RA_CW_CNT feature; and for the single hypothesis facet with the lowest entailment probability, the POS tags of its governor and modifier and their aligned terms from the text. We then trained a support vector machine on the RTE feature vectors from RTE3 development and RTE1 and RTE2 data and used it to classify the RTE3 test data. The results are discussed in the following section.

4.8 Results on RTE Challenge Data

Table 6 shows the accuracy on the RTE3 test data along with accuracy in 10-fold cross validation on the training data. The first column provides the majority class (Entailed) baseline. The second column presents results considered to be a reasonable string comparison baseline (c.f., Giampiccolo, Magnini, Dagan and Dolan 2007), which were also the median results across all RTE3 challenge submissions. The third column shows the accuracy when training strictly on the facet entailment probability estimates output by the ASK corpus trained classifier. The final column presents accuracy using all of the features described in the previous section. Examining per class performance reveals that the classifier is much more accurate on entailed pairs (84.1%) than non-entailed pairs (49.2%). These results would have been fourth of twenty-three in the RTE3 challenge, but as with all others, fall significantly short of the best results (80%) achieved by Hickl and Bensley (2007).

Table 6. Classifier accuracy on RTE3 test set

	Majority Label	String Matching Baseline	Facet Probability Estimates Only	All Features
Training Set 10x CV	50.3	n/a	58.4	62.6
RTE3 Test Set	51.3	61	63.8	67.1

One of the primary goals of our paradigm within the ITS domain is to facilitate a dialogue focused on the specific facet of the reference answer where the student's response could be improved. Along those lines, perhaps the more useful application to the RTE data would be to identify why and in what area a hypothesis is not entailed. While this is largely left as future work, the RTE facet accuracy presented in the previous section suggests the approach should be reasonably successful.

5 Conclusion

The results on the children's corpus presented here are 19.4% over a lexical baseline derived from the best performing system at the first RTE challenge. This demonstrates that the task is feasible and we believe with more rigorous feature engineering, accuracy should be in a range that allows effective tutoring. We are currently in the process of implementing our ITS and a chief area of future research is incorporating this system and assessing its effectiveness with spoken dialog input. In prior work (Nielsen, Ward and Martin 2007), we achieved confidence weighted scores approximately 10% (absolute) over the classification accuracy, indicating that the class probability estimates will be useful to the dialog manager in deciding how strongly to believe in the classifier's output. If the classification suggests the learner understood a concept, but the confidence is low, the dialog manager could paraphrase the answer as a transition to the next question, rather than assuming the learner definitely understands and simply moving on and rather than asking a confirming question about something the learner probably already expressed. Thus, even if there is a moderate error rate in classifying the facets, confidence estimates can still facilitate effective dialogue.

Three of the most significant contributions of this work are defining and evaluating a new paradigm for learner answer assessment that involves classification of fine grained answer facets with the detailed labels necessary to enable more intelligent dialog control, creating an annotated corpus that supports this paradigm, and laying the framework for an answer assessment system that can classify learner responses to previously unseen questions according to this scheme. This is also the first work to demonstrate success in assessing short constructed responses from elementary school children.

The validity of the entailment paradigm, semantic representation, and corpus annotation described here is strengthened by substantial interannotator agreement (86.2%, Kappa = 0.728) and by promising automated classification results. The corpus is available for researchers to utilize in improving other entailment systems, as well as tutoring and educational assessment technologies. This database of annotated answers provides a shared resource and a standardized annotation scheme that we hope will stimulate further research in these areas.

All prior work on intelligent tutoring systems has focused on question specific assessment of answers and even then the understanding of learner responses has generally been restricted to a judgment regarding their correctness or in a small number of cases a classification that specifies which of a predefined set of misconceptions the learner might be exhibiting. The domain independent approach described here enables systems that can easily scale up to new content and learning environments, avoiding the need for lesson planners or technologists to create extensive new rules or classifiers for each new question the system must handle. This is an obligatory first step in creating intelligent tutoring systems that can truly engage children in natural unrestricted dialog, such as is required to perform high quality student directed Socratic tutoring.

References

Agichtein E, Gravano L (2000) Snowball: Extracting relations from large plain-text collections. In Proc. of the 5th ACM ICDL

Aleven V, Popescu O, Koedinger KR (2001) A tutorial dialogue system with knowledge-based understanding and classification of student explanations. In IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems

Bar-Haim R, Szpektor I and Glickman O (2005) Definition and Analysis of Intermediate Entailment Levels. In *Proc. Workshop on Empirical Modeling of Semantic Equivalence and Entailment*.

Barzilay R, Lee L (2003) Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In Proc. of HLT-NAACL, pp 16–23

Barzilay R, McKeown K (2001) Extracting paraphrases from a parallel corpus. In Proc. of the ACL/EACL, pp 50–57

Braz RS, Girju R, Punyakanok V, Roth D, Sammons M (2005) An inference model for semantic entailment in natural language. In Proc. of the PASCAL Recognizing Textual Entailment challenge workshop

Burger J, Ferro L (2005) Generating an entailment corpus from news headlines. In Proc. of the ACL workshop on Empirical Modeling of Semantic Equivalence and Entailment, pp 49–54

Callear D, Jerrams-Smith J, Soh V (2001) CAA of short non-MCQ answers. In Proc. of the 5th International CAA conference, Loughborough

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

Dagan I, Glickman O, Magnini B (2005) The PASCAL Recognizing Textual Entailment Challenge. In Proc. of the PASCAL RTE challenge workshop

Dolan WB, Quirk C, Brockett C (2004) Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. Proceedings of COLING 2004, Geneva, Switzerland.

Giampiccolo D, Magnini B, Dagan I and Dolan B. (2007) The Third PASCAL Recognizing Textual Entailment Challenge. In Proc. of the ACL-PASCAL workshop on textual entailment and paraphrasing.

Gildea, D. & Jurafsky, D. (2002). Automatic labeling of semantic roles. Computational Linguistics, 28:3, 245–288.

Glickman O, Dagan I (2003) Identifying lexical paraphrases from a single corpus: A case study for verbs. In Proc. of RANLP

Glickman O, Dagan I, Koppel M (2005) Web based probabilistic textual entailment. In Proc. of the PASCAL RTE challenge workshop

Graesser AC, Hu X, Susarla S, Harter D, Person NK, Louwerse M, Olde B, the Tutoring Research Group (2001) AutoTutor: An intelligent tutor and conversational tutoring scaffold. In Proc. of the 10th International Conference of Artificial Intelligence in Education. San Antonio, TX, pp 47–49

Hickl A, Bensley J (2007) A discourse commitment-based framework for recognizing textual entailment. In Proc. of the ACL-PASCAL workshop on textual entailment and paraphrasing.

Kipper K, Dang HT, Palmer M (2000) Class-based construction of a verb lexicon. AAAI seventeenth National Conference on Artificial Intelligence

Landauer TK, Dumais ST (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. J Psychological Review

Lawrence Hall of Science (2005) Full Option Science System (FOSS), University of California at Berkeley, Delta Education, Nashua, NH

Lawrence Hall of Science (2006) Assessing Science Knowledge (ASK), University of California at Berkeley, NSF-0242510

Leacock C, Chodorow M (2003) C-rater: Automated scoring of short-answer questions. Computers and the Humanities, 37:4

Lin D, Pantel P (2001) Discovery of inference rules for question answering. In Natural Language Engineering, 7(4):343-360

MacCartney B, Grenager T, de Marneffe M, Cer D, Manning C (2006) Learning to recognize features of valid textual entailments. In Proc. of HLT-NAACL

Makatchev M, Jordan P, VanLehn K (2004) Abductive theorem proving for analyzing student explanations and guiding feedback in intelligent tutoring systems. Journal of Automated Reasoning special issue on automated reasoning and theorem proving in education. 32(3):187-226

Mitchell T, Russell T, Broomhead P, Aldridge N (2002) Towards robust computerized marking of free-text responses. In Proc. of 6th International Computer Aided Assessment Conference, Loughborough

Nielsen RD, Ward W, Martin JH (2006) Toward dependency path based entailment. In Proc. of the second PASCAL RTE challenge workshop

Nielsen RD, Ward W (2007) A corpus of fine-grained entailment relations. In Proc. of the ACL workshop on Textual Entailment and Paraphrasing

Nielsen RD, Ward W, Martin JH (2007) Soft computing in Intelligent Tutoring Systems and Educational Assessment. In Soft Computing Applications in Business. Springer

Nivre J, Hall J, Nilsson J, Chanev A, Eryigit G, Kubler S, Marinov S, Marsi E (2007) MaltParser: A language-independent system for data-driven dependency parsing. Natural Language Engineering, 13(2):95-135

Pang B, Knight K, Marcu D (2003) Syntax-based alignment of multiple translations: Extracting paraphrases and generating sentences. In Proc. HLT/NAACL

Pon-Barry H, Clark B, Schultz K, Bratt EO, Peters S (2004) Contextualizing learning in a reflective conversational tutor. In Proceedings of the 4th IEEE International Conference on Advanced Learning Technologies

Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann.

Raina R, Haghighi A, Cox C, Finkel J, Michels J, Toutanova K, MacCartney B, de Marneffe MC, Manning CD, Ng AY (2005) Robust textual inference using diverse knowledge sources. In Proc of the PASCAL RTE challenge workshop

Ravichandran D, Hovy E (2002) Learning surface text patterns for a question answering system. In Proc. of the 40th ACL conference. Philadelphia

Rosé CP, Roque A, Bhembe D, VanLehn K (2003) A hybrid text classification approach for analysis of student essays. In Building Educational Applications Using Natural Language Processing, pp 68-75

Sudo K, Sekine S, Grishman R (2001) Automatic pattern acquisition for Japanese information extraction. In Proc. of HLT, San Diego, CA

Sukkarieh JZ, Pulman SG, Raikes N (2003) Auto-marking: Using computational linguistics to score short, free text responses. In Proc. of the 29th conference of the International Association for Educational Assessment. Manchester, UK

Tatu M, Moldovan D (2007) COGEX at RTE 3. In Proc. of the ACL-PASCAL workshop on Textual Entailment and Paraphrasing

Turney PD (2001) Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In Proc. of 12th European Conference on Machine Learning, pp 491–502

Vanderwende L, Coughlin D and Dolan WB (2005) What Syntax can Contribute in the Entailment Task. In *Proc. of the PASCAL Workshop for Recognizing Textual Entailment*.