LEARNER ANSWER ASSESSMENT
IN INTELLIGENT TUTORING SYSTEMS

by

RODNEY D. NIELSEN

M.S., University of Colorado, Boulder, 2005

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Computer Science

Institute of Cognitive Science

2007

This thesis entitled:
Learner Answer Assessment in Intelligent Tutoring Systems
written by Rodney D. Nielsen
has been approved for the Department of Computer Science
and Institute of Cognitive Science

_____
Wayne Ward

_____
James Martin

Date_____

The final copy of this thesis has been examined by the signatories, and we
Find that both the content and the form meet acceptable presentation
standards
Of scholarly work in the above mentioned discipline.

Nielsen, Rodney D. (Ph.D., Computer Science and Cognitive Science)

Learner Answer Assessment in Intelligent Tutoring Systems

Thesis directed by Professors Wayne H. Ward and James H. Martin

Abstract

Truly effective dialog and pedagogy in Intelligent Tutoring Systems will only be achievable when systems are able to understand the detailed relationships between a learner's answer and the desired conceptual understanding. This thesis describes a new paradigm and framework for recognizing whether a learner's response to an automated tutor's question entails that they understand the concepts being taught. I illustrate the need for a finer-grained analysis of answers than is supported by current tutoring systems and describe a new representation for reference answers that addresses these issues, breaking them into detailed facets and annotating their relationships to the learner's answer more precisely. Human annotation at this detailed level still results in substantial inter-annotator agreement, 86.1%, with a Kappa statistic of 0.728.

I present current efforts to automatically assess learner answers within this new framework, which involves training machine learning classifiers on features extracted from dependency parses of the reference answer and the learner's response and features derived from domain-independent lexical statistics. The system's performance, 75.5 % accuracy within domain and 65.9% out of domain, is very encouraging and confirms the approach is feasible.

Another significant contribution of this work is that the semantic assessment of answers is completely domain-independent. 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 handcraft information extraction templates.

Contents

List of Tables

List of Figures

1 Introduction

Imagine a time in the future when, in addition to instructor-led group interaction in the classroom, children get one-on-one or small group (two or three to one) tutoring with a subject matter expert. At a time when funding for education seems to be continually cut, this seems impossible to imagine and almost surely never will come to be, at least not with human tutors, but what about computers? Might it be possible for computers to engage students in this same form of natural face-to-face conversation? Might they even be more capable of patiently tailoring their interactions to each child's learning style or knowing just the right question to ask at just the right time to maximize learning outcomes? What capabilities must the system possess to carry out this feat? While there are many unsolved problems between today and the future these questions evoke, this thesis takes steps in the direction of solving one of these problems – getting a machine to understand a child's utterance in the context of a tutor's question.

Consider the question in example (1a), the desired answer in (2a), and the child's answer shown in (3a).

(1a) *Kate dijo: "Un objeto tiene que moverse para producir sonido." ¿Estás de acuerdo con ella? ¿Por qué sí o por qué no?*

(2a) *De acuerdo. Las vibraciones son movimientos y las vibraciones producen sonido.*

(3a) *Sí, porque los sonidos vibran, chocan con el objeto y se mueve.*

What is required for the computer to understand the relationship between the student's answer and the reference answer? The relationship seems obvious to us, well, at least if we understand the language, but what would we do if, like the computer, we were not fluent in the language? One approach utilized by many is to examine the lexical similarity between the answers. We might recognize in this case that the student used other derivational forms of nearly all of the right words and hence, assume that they answered the question correctly. Unfortunately, the problem is not that easy. In fact, in this case, the student got the cause and effect completely backwards, as can be seen in the original English versions (1b), (2b) and (3b). The student was on topic, used virtually all the right words, but clearly does not understand the concepts involved. It is critical that the computer be able to assess the interplay and relations between the words.

(1b) *Kate said: "An object has to move to produce sound." Do you agree with her? Why or why not?*

(2b) *Agree. Vibrations are movements and vibrations produce sound.*

(3b) *Yes, because sounds vibrate and hit the object and it moves.*

The overarching thesis of this work is that a more detailed assessment of learners' dialog contributions will enable tutoring strategies that will significantly improve learner comprehension. The thesis that this work more directly addresses is that, with the use of fully automated systems, learner contributions can be classified at a level assumed to be appropriate for achieving the above goal and that this assessment can be performed in a domain-independent manner. Such a level of analysis would have to meet the following criteria:

- It must utilize a level of representation that facilitates a detailed assessment of the learner's understanding, indicating exactly where and in what manner the answer did not meet expectations.
- The representation and assessment must be learnable by an automated system – they must not require the handcrafting of domain-specific logic representations, parsers, knowledge-based ontologies, or dialog management rules.

This thesis presents just such a paradigm shift with a new framework for assessing learner responses to tutor questions. I break the reference answer down into very low-level compositional facets and annotate their relationships to the student's answer more precisely than has been done in prior work. I describe an initial approach to automatically assess answers within this framework with the long term goal of improving the state of intelligent tutoring systems to the point where they are comparable to or can even outperform human tutors, as measured by increases in the associated student learning gains. In order to achieve this long-term goal I believe the tutor must be capable of natural, engaging, domain-independent dialog.

Another significant contribution of this work is that the semantic assessment of answers is domain-independent – the system does not need to be retrained for new questions or even new subject areas. No prior work in the area of tutoring systems or answer verification has attempted to build such question-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 handcraft information extraction templates. Because comprehension problems often take root in elementary school during the early years of learning to read and comprehend texts, this thesis focuses on those critical grades. To my knowledge, this is the first work to show success in assessing elementary students' roughly sentence-length responses to comprehension questions.

I begin with a brief motivation for why intelligent tutoring systems (ITSs) are an important component of future educational settings. Then I describe early, more conventional approaches to assess students' conceptual understanding in automated tutoring systems along with their shortcomings. I describe the progression of strategies and discuss how they improved over earlier efforts and indicate where they still have room for improvement. Technological approaches to large-scale assessment are then reviewed and contrasted with automated tutoring technology. I also describe the relevance of some current active areas of natural language processing research such as paraphrase recognition and textual entailment. I then detail my semantic analysis framework, the generation of a gold standard annotated corpus within this framework, and my current efforts to achieve more robust assessment of understanding by applying machine learning to detect the relationships between phrases that entail an understanding of a tutored concept and those that do not.

2 Why are Intelligent Tutoring Systems Important?

Improving reading comprehension is a national priority, particularly in the area of science education. In 1999, the RAND Reading Study Group (Snow 2002), which was commissioned by the Department of Education to develop a national agenda for reading research, concluded that "Understanding how to improve reading comprehension outcomes, not just for students who are failing in the later grades but for all students who are facing increasing academic challenges, should be the primary motivating factor in any future literacy research agenda."

Recent estimates suggest that over one third of fourth grade readers and 27% of 8th grade readers cannot extract the general meaning nor even make simple inferences from grade-level text (National Assessment of Educational Progress, NAEP 2007); the most recent results for 12th grade readers suggest that 27% of them also fall within this category (NAEP 2005). While many students may appear to learn to read and understand text by third grade, evidence shows that their apparent competence is often an illusion – as texts become more challenging in fourth grade, many students cannot read nor understand them (Meichenbaum and Biemiller 1998; Sweet and Snow 2003). There is thus a critical need for programs that engage beginning readers in a way that supports comprehension.

The lack of sufficient comprehension of texts is a significant contributing factor to poor learning outcomes in science (Gomez et al. in press), leading to the current state of U.S. science literacy: "Current levels of mathematics and science

achievement at the elementary and secondary levels suggest that the United States is neither preparing the general population with levels of mathematics and science knowledge necessary for the 21st century workplace, nor producing an adequate pipeline to meet national needs for domestic scientists" (The Institute of Educational Sciences 2006). In the most recent National Assessment of Educational Progress in science (NAEP 2005), only three percent of U.S. students attained advanced levels of science achievement in Grades 4 and 8, with even fewer reaching advanced levels in Grade 12. Many U.S. students are not even attaining mastery of rudimentary science knowledge and skills. In the 2005 NAEP, 32 percent of Grade 4 students, 41 percent of Grade 8 students, and 46 percent of Grade 12 students scored below the *Basic* level in science. At Grade 4, students performing below the basic level cannot read simple graphs. At Grade 12, students performing below the basic level are likely to miss problems such as drawing a simple diagram of the solar system. Only 29 percent (corrected for chance) of all Grade 4 students recognized that the moon's craters are the result of meteoroid impacts versus eruptions of active volcanoes, shifting rock (moonquakes), or tidal forces caused by the Earth and Sun. Only 62 percent were able to recognize and explain why, when dropped in identical glasses of water, the significantly larger steel ball in the problem's drawing would cause a greater rise in the water level.

How do we tailor instruction to accommodate differing learners' needs and address these comprehension deficits? Comprehension of text can be facilitated by reading texts to kids and then interacting with them in various ways

to promote thinking and learning (e.g., Beck and McKeown 2001; Chi 1996; VanLehn et al. 2003). One of the long term goals associated with this work is to develop an intelligent tutoring system that both facilitates the comprehension of a given text and teaches students to consider critical questions and thus form a deeper understanding when reading future texts.

Since accommodating differing learner needs through tailored instruction is at the core of the project's long term goals, the mode of teaching and learning on which this thesis is based is one-on-one dialog, the kind of instruction that Benjamin Bloom observed so greatly impacts learning and cognition. In 1984, Bloom determined that the difference between the amount and quality of learning for a class of thirty students and those who received individualized tutoring was 2 standard deviations. The significant differences in proficiency between those children who enjoy one-on-one tutoring versus those who have little or no individualized support is testament to the need for further exploration of the individualized tutoring model (Bloom 1984; Torgesen, Wagner and Rashotte 1999).

In the two decades since Bloom reported a two sigma advantage of one-on-one tutoring over classroom instruction across several subjects, evidence that tutoring works has been obtained from dozens of well designed research studies, meta-analyses of research studies (e.g., Cohen, Kulik and Kulik 1982) and positive outcomes obtained in large scale tutoring programs in Britain (e.g., Topping and Whitley 1990) and the U.S. (Madden and Slavin 1989). Effective intelligent tutoring systems that produce learning gains with high school, college,

and adult subjects through text-based dialog interaction exist in the laboratory (e.g., Graesser et al. 2001; Peters et al. 2004; VanLehn et al. 2005), some demonstrating up to a one sigma gain relative to classroom instruction (Anderson et al. 1995; Koedinger et al. 1997). Therefore, intelligent tutoring systems show promise in approaching the effectiveness of human tutoring and systems that are accessible, inexpensive, scalable and above all effective would provide one critical component of an overall educational solution.

From a cost-benefit perspective, computers and associated learning software have the potential to provide a relatively inexpensive solution in today's education system.

The cost of training and employing additional teachers to provide the level of individualized attention that many students need is prohibitively expensive, whereas interactive computer systems that use advanced human communication and interface technologies to provide individualized tutoring could inexpensively provide focused, individualized, adaptive, scientifically-based instruction. In short, advances in computing technologies, communications, and language technologies combined with advances in cognitive science and the science of reading and learning, provide a powerful and timely potential solution to our nation's education crises.

3 Prior Work on Intelligent Tutoring Systems

Early automated tutors largely followed conventional computing approaches, for example, utilizing multiple choice questions. Current research shows that, if properly designed, multiple choice questions can be a very effective tool for assessing comprehension. However, one of the most successful means of improving learning gains is to force students to articulate their beliefs, leading them to a better understanding of what they do and do not know and strengthening the causal relations between the bits of knowledge they have acquired (Chi 1996; VanLehn et al. 2003). Therefore, much research has moved away from multiple choice questions and toward free response questions, requiring students to express their deeper understanding of the concepts in the text.

The first most obvious step in this direction was the use of scripted, domain-specific dialog techniques, often implemented utilizing Finite State Machines (FSM). For example, SCoT, a Spoken Conversational Tutor (Peters et al. 2004), uses *activity recipes* to specify what information is available to the recipe, which parts of the information state are used in determining how to execute the recipe, and how to decompose an activity into other activities and low-level actions. The Phoenix semantic parser (Ward 1991) defines semantic frames which include patterns to be matched and extracted from the dialog. The dialog is then driven by what frame elements have not yet been addressed and the specific system turns attached to those frame elements. The advantage of this

technology lies in its precision[1] – it is generally very accurate in the classification of relevant text fragments, but this is typically at the cost of poorer performance on recall[2] – finding all of the relevant information nuggets.

We conducted a pilot experiment to determine how well an automatic system based on domain-specific shallow semantic parsing (the process of recognizing predicate-argument structure or semantic relationships in text) via the Phoenix parser could grade young children's summaries of stories. We collected spoken summaries of a single story, "Racer the Dog", from 22 third and fourth grade students. We divided the summaries into a training set consisting of fifteen summaries and a test set of seven summaries. Multiple researchers generated reference summaries, which we distilled to consist of the key points that should exist in a good summary of the story. We utilized the Phoenix semantic parser to map summaries to semantic frames (see Fig. 1 and Fig. 2 for examples).[3] We first wrote an initial parser grammar that could extract the points in the reference summary and then utilized the fifteen training summaries to expand the coverage of the grammar. The seven summaries comprising the test set were not examined during system development. After system development, we manually parsed the

---

[1] Precision is calculated as the fraction of classifications that are correct out of all of the text fragments the system classified as falling within one of the categories of interest.

[2] Recall is the fraction of classifications that are correct out of all of the text fragments that should have been classified as falling within one of the categories of interest according to the human gold standard annotation.

[3] Anaphoric reference was resolved manually in a preprocessing step.

test set to create gold standard reference parses for evaluation purposes. The manual parse identified a total of 36 points that addressed content from the reference summaries. Compared to this gold standard, automatic parses of the test set had a recall of 97% (35/36) and a precision of 100% (35/35) – the parser found all but one of the relevant points and produced no erroneous ones. (In order to consider a parse correct, the concept from the child's summary had to have the same semantic roles as one in the reference summary and the roles had to be filled by references to the same entities.)

Racer had problems with his back legs.

LegProblems:[Problems_agent].[Dog_Name].racer

Fang always bit Racer and ran away faster than Racer could run.

Bother:[Bother_agent].[Dog_Name].fang

Bother:[Bother_theme].[Dog_Name].racer

Faster:[Run_Away_agent].[Dog_Name].fang

Faster:[Run_Away_theme].[Dog_Name].racer

Fig. 1. Example Phoenix parse for a reference summary

um racer got his leg hurt

LegProblems:[Problems_agent].[Dog_Name].racer

fang kept on teasing racer

Bother:[Bother_agent].[Dog_Name].fang

Bother:[Bother_theme].[Dog_Name].racer

Fig. 2. Example Phoenix parse of a child's summary

The preceding results are for human transcriptions of the children's spoken summaries. We also decoded the speech files with the University of Colorado SONIC speech recognition system (Pellom 2001; Pellom and Hacioglu 2003) and processed the SONIC output. For the same test set, the recall was 83% (30/36) and the precision was 100% (30/30) – speech recognition errors resulted in missing an additional five points that were extracted from the human transcribed summaries, but generated no erroneous points. As expected by such systems, the precision was quite high, while the recall was slightly less so.

A significant disadvantage of these more conventional systems is that they require a considerable investment in labor to cope with a new subject area or even to handle a small change in subject matter coverage. This effort is required to generate new handcrafted parsers, knowledge-based ontologies, and dialog control mechanisms. In this regard, the use of Latent Semantic Analysis (LSA), a statistical soft computing technique, to assess student's summaries represents an improvement over FSM dialogs, in that the system is more flexible in handling the unconstrained responses of a learner (Landauer and Dumais 1997; Landauer, Foltz and Laham 1998). LSA begins with a term by document matrix, where the cells in the matrix indicate the number of occurrences of the given term in the associated document. This matrix is given a TF-IDF (term frequency – inverse document frequency) weighting to account for the relative importance of a term in the document adjusted for its significance across all documents. LSA then utilizes singular value decomposition and retains only the top k (usually around 300) dimensions to represent the key information in the original matrix. This

process effectively smoothes the data and brings out the latent semantics in the original document set – providing connections between terms and documents that were not explicit in the source text. The resulting rank-*k* approximation of the original matrix is then used to determine the similarity of terms and documents by calculating the dot product or cosine with related vectors.

The Institute of Cognitive Science (ICS) and The Center for Spoken Language Research (CSLR), both at the University of Colorado, Boulder, have worked with the Colorado Literacy Tutor program to develop educational software that helps children learn to read and comprehend text (Cole et al. 2003; Franzke et al. 2005). A significant part of this program is Summary Street, a tool for improving and training text comprehension through summarization. Summary Street utilizes LSA to grade children's text summaries and provide feedback on the quality of the summary, including completeness, relevance and redundancy. Summary Street's feedback has been shown to improve student scores by, on average, a letter grade with much more improvement for lower performing students.

AutoTutor (Mathews et al. 2003) is an interactive text-based tutor that utilizes LSA to engage college students in dialogs regarding conceptual physics and introductory computer science. Because AutoTutor is based on LSA, it is more flexible in handling the unconstrained responses of a learner than are tutors based on FSM dialogs. The AutoTutor architecture requires the lesson planner or system designer to provide the following information:

1. A statement of the problem to be solved, in the form of a question.
2. A set of expectations in an ideal answer, with each expectation being a sentence in natural language of 10-20 words
3. A set of tutor dialog moves that express or elicit from the learner each expectation in #2 (i.e., hints, prompts, and assertions)
4. A set of anticipated bad answers and corrections for those bad answers
5. A set of [subject matter] misconceptions and corrections to those misconceptions
6. A set of basic noun-like concepts about [the subject matter] and their functional synonyms in the specific context of the problem.
7. A summary of the answer or solution
8. A latent semantic analysis (LSA) vector for each expectation, bad answer, and misconception.

(Mathews et al. 2003)

The key advantage of their architecture is that it does not require a syntactic match between the learner's dialog turn and the expectation in the ideal answer; nor does it require key phrase spotting. The LSA component analyzes the semantic similarity between the user's open-ended dialog turn and the system's expectation or reference answer by comparing vector representations derived from a bag-of-words method – a method which ignores the expressed relations between words. The dialog is then driven by checking the degree to which each expectation has been covered by the sum total of all past user turns for the problem. The given answer expectation and the compilation of user turns are each represented as pseudo-documents, weighted averages of the reduced-dimensionality vectors associated with the words they contain. The answer is then assessed by computing the cosine between the vectors representing the two

pseudo-documents. If the metric exceeds a threshold, the user's turns are assumed to have the same meaning as the answer expectation.

However, the reality is that, while LSA's evaluations are closely correlated with human evaluations, LSA completely disregards syntax in its analysis and is prone to many related weaknesses. In example (3) above, (*sounds vibrate and hit the object and it moves*), the student used all the right key words, so LSA would be satisfied despite the fact that the student has the causal relation reversed. LSA does a poor job of detecting misconceptions (Mathews et al. 2003); it seems likely that this might be due to the relatedness between the bag of words in the misconception's description and the bag of words in the expected answer to the question. LSA also performs very poorly on the short answers that are typical in tutoring settings. This is the reason that AutoTutor must combine the input from all prior user turns into one cumulative bag of words, rather than compare strictly with the learner's current response. Furthermore, and key to this thesis, it is not possible with typical LSA-based approaches to perform a detailed assessment of a learner's contribution or to identify the specific reasons that a short answer might not be correct. Consequently, typical LSA approaches provide little help in classifying learner contributions and determining the best tutor response or dialog strategy to correct misconceptions. An additional goal of the work presented here is to develop a system that does not need to be retrained for each new question or subject area.

While the LSA-based approaches are not typically trained for individual questions, they do generally require a fair amount of corpus tuning to ensure

adequate coverage of the topic area and the cosine threshold to judge similarity is always tuned to the domain or question. As noted above, AutoTutor also requires "a set of basic noun-like concepts about [the subject matter] and their functional synonyms in the specific context of the problem" (Mathews et al. 2003). Lastly, there is no evidence that LSA is an effective tool for interacting with young children in the K-6 grade range. In preliminary investigations using story summaries written by third and fourth graders, we found that LSA was unable to appropriately assess the quality of these children's summary-length responses. This is further confirmed by LSA's poor performance in grading elementary students' constructed responses to short answer questions in experiments by the creators of AutoTutor (Graesser, personal communication).

Rosé et al. (2003a) and Jordan, Makatchev and VanLehn (2004) have improved on the accuracy of the LSA-based approach to verifying short answers in tutorial dialogs by incorporating a deep syntactic analysis into their evaluation and integrating multiple assessment technologies. A decision tree is used to learn from the output of a Naïve Bayes classifier and from deep syntactic parse features.[4]

The system classifies each student sentence to determine which, if any, of the set of good answer aspects it matches. This is one of the very few systems for assessing open-ended short answers that provides any finer-grained measure of performance than a simple answer grade; they evaluate the system using

---

[4] Their deep parse includes information beyond a typical syntactic parse (e.g., mood, tense, negation, etc.)

precision, recall, and F-scores on the task of detecting the good answer aspects in the student responses. They compare their combined technique, CarmelTC, to LSA as a baseline, as well as to the Naïve Bayes classifier without the syntactic parse features, and a classifier based only on the syntactic features without the Naïve Bayes output. The results, as seen in Table 1, show that the combined approach performed much better than LSA or either of the individual systems.

Method	Precision	Recall	F
LSA	93%	54%	0.70
Naïve Bayes	81%	73%	0.77
Symbolic-only	88%	72%	0.79
CarmelTC	90%	80%	0.85

Table 1. Evaluation of CarmelTC

Though CarmelTC does provide slightly more information than a simple grading system, it does not provide the sort of fine-grained assessment of a learner's contribution necessary to understand their mental model and drive high-quality tutoring dialog. Furthermore, much of the parsing in their system is dependent on domain-specific, hand-coded rules, in order to capture the semantics of the domain lexicon and language. While they are building tools to ease this process (Rosé et al. 2003b), there appears to be a long way to go before a lesson planner, with no formal linguistics or computer science background, would see a positive cost-benefit analysis in utilizing these tools to build lessons, especially on a regular basis. For example, to handle a new question the user must generate first order propositional representations of each good answer aspect and hand-

annotate example learner answers as to their semantic interpretation within those representations.

Perhaps more important than the time-consuming nature of this process, it defeats one of the ultimate goals of an Intelligent Tutoring System (ITS), namely to be able to interact naturally without being constrained to communication within the realm of a fixed set of question-answer pairs. Finally, the methods described by Rosé and Jordan require that classifiers be trained for each possible proposition associated with the reference answers. This not only requires additional effort (currently on the part of the systems builders) to train the system, but it also requires the collection and grading of a large corpus of learner responses to each question that the tutor might ask; again defeating the ultimate goal of natural dynamic tutoring interactions.

Makatchev, Jordan and VanLehn (2004) develop an abductive theorem prover with the shared long-term goal of providing more specific, higher quality feedback to learners. However, their approach is domain dependent, requiring extensive knowledge engineering for each new qualitative physics problem to be solved; their reasoning engine has 105 domain rules that handle seven specific physics problems.

Recall and precision vary by proof cost; at a proof cost of 0.6, recall and precision are approximately 0.38 and 0.25, respectively, where at a proof cost of 0.2, they are around 0.63 and 0.15, leading to F-measures of about 0.30 and 0.24 respectively. In initial analyses, their system did not statistically improve learning outcomes above those achieved by having the students simply read the text. They

also indicate that the system performed poorly at correctly identifying misconceptions in the student essays.

Many other ITS researchers are also striving to provide more refined learner feedback (e.g., Aleven, Popescu, and Koedinger 2001; Peters et al. 2004; Pon-Barry et al. 2004; Roll et al. 2005). However, they too are developing very domain-dependent approaches, requiring a significant investment in handcrafted 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 never enable the long-term goal of effective unconstrained interaction with learners or the pedagogy that requires it.

4 Related Research

4.1 Short Constructed-Response Scoring

There is a small, but growing, body of research in the area of scoring free-text responses to short answer questions (e.g., Boonthum 2004; Callear, Jerrams-Smith and Soh 2001; Leacock 2004; Leacock and Chodorow 2003; Mitchell et al. 2002; Mitchell, Aldridge and Broomhead 2003; Pulman 2005; Sukkarieh, Pulman and Raikes 2003; Sukkarieh and Pulman 2005). Shaw (2004) and Whittington (1999) provide reviews of some of these approaches. Most of the systems that have been implemented and tested are based on Information Extraction (IE) techniques (Cowie and Lehnert 1996). They handcraft a large number of pattern rules, directed at detecting the key aspects of correct answers or common incorrect answers. When used for high school and college age students, the results of these approaches range from about 84% accuracy up to nearly 100%. However, the implemented systems are nearly all written by private companies that keep much of the nature of the questions and systems proprietary and the best results seem to frequently be achieved by tuning with the test data. Still, these results provide good evidence that answers can be accurately assessed, at least at the coarse-grained level of assigning a score to high school and college level students' answers.

Work in this area is typified by c-rater (Leacock and Chodorow 2003; Leacock 2004), which was designed for large-scale tests administered by the Educational Testing Service (ETS). A user examines 100 example student

answers and manually extracts the common syntactic variants of the answer (200 or more examples are used for problematic questions that have more syntactic variation or fewer correct answers in the dataset, etc). They build a model for each syntactic variant and specify which aspects of the subject-verb-object structure are required to give credit for the answer. For each of the subject, verb, and object, the model authoring tool provides the user with a list of potential synonyms and they select those synonyms that are appropriate within the given context. The list of synonyms is extracted via the information theoretic similarity metrics described in Lin (1998). Lin's similarity metric computes the mutual information between words that are connected via syntactic dependencies and then uses this to calculate the similarity between any pair of words based on the ratio of information in the dependencies the two words have in common to the sum of information in all dependencies involving either word. The more dependencies two words have in common, the closer this ratio is to 1.0. Given the list of similar words, the user of c-rater then selects those synonyms that are appropriate within the given context.

During the process of scoring a student's answer, the verb lemma is automatically determined and it must match a lemma in one of the model answers; the system also resolves pronouns and attempts to correct non-word misspellings. Scoring consists of assigning a value of 0 (no credit), 1 (partial credit), or 2 (full credit). In a large-scale reading comprehension exam administered across Indiana (16,625 students), c-rater achieved an 84% accuracy

as measured relative to the scores of human judges on a random sample of 100 answers for each of the seven questions that they were able to score using c-rater.

In general, short constructed-response scoring systems are designed for large scale assessment tasks, such as those associated with the tests administered by ETS. Therefore, they are not designed with the goal of accommodating dynamically generated, previously unseen questions. Similarly, these systems do not provide feedback regarding the specific aspects of answers that are correct or incorrect; they merely provide a raw score for each question. As with the related work directed specifically at ITSs, these approaches all require in the range of 100-500 example student answers for each planned test question to assist in the creation of IE patterns or to train a machine learning algorithm used within some component of their solution.

4.2 Paraphrasing and Entailment

In recent years, there has been a tremendous increase in interest in the areas of paraphrase acquisition and textual entailment recognition or proof. These technologies have broad applications in numerous areas and great relevance to this work. Paraphrasing is the most common means for a learner to express a correct answer in an alternative form; in fact, 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 true entailments.

The target of a fair amount of the work on paraphrasing is in acquiring paraphrases to be used in information extraction or closely related factoid question answering (Agichtein and Gravano 2000; Duclaye, Yvon and Collin 2002; Ravichandran and Hovy 2002; Shinyama and Sekine 2003; Shinyama et al. 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). For example, Agichtein and Gravano (2000) use a bootstrapping approach to automatically learn paraphrase patterns from text, given just five seed examples of the desired relation. The technique does not identify general patterns, just those associated with the relation in the seeds. Beginning with the seed examples they extract patterns in the form of a five-tuple <left-context, entity-A, middle-context, entity-B, right-context>, where the contexts are represented in a soft form as weighted vectors that disregard word order. The terms in each context are weighted according to their frequency across the seeds. These patterns are used to find additional entity pairs assumed to have the same relation, which are then used as seed examples to find additional patterns and the process repeats. They achieved almost 80% recall and 85% precision in the task of extracting headquarters' locations for organizations. This could be applied to answer assessment, if you can automatically identify the "entities" in an answer using Named Entity (NE) recognizers, then via a similar technique you can determine whether the two encompassing contexts are essentially paraphrases. Since fact-based questions are also somewhat common in testing environments, these paraphrase extraction techniques could be useful in tutoring systems. However, I

am more interested in questions that promote deeper reasoning than simple fact 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. From these alignments, Barzilay and McKeown extract patterns for valid lexical or short phrasal paraphrases. They use a machine learning algorithm that is a variant of Co-Training based on (a) context features and (b) lexical and POS features. First, they initialize the seed set of paraphrases to match identical word sequences in aligned sentences, then they iteratively find contexts based on the paraphrases and paraphrases based on the contexts, until no new paraphrases are found or a specified number of iterations passes. They found 25 morpho-syntactic rules and 9,483 paraphrases, of which about 29% were multi-word phrases. Judges found the paraphrases (with context provided) to be valid about 91.6% of the time. Only around 35% of the lexical paraphrases extracted by these techniques were synonyms; the remaining were 32% hypernyms, 18% siblings of a hypernym, 5% from other WordNet relations, and 10% were not related in WordNet, illuminating the need for softer assessment in tutoring systems versus requiring strict synonyms.

Pang, Knight and Marcu assess syntactic constituency trees to generate Finite State Automata that represent paraphrases. Barzilay and Lee also generate word lattices, but from a single corpus by combining the surface forms of several similarly written sentences about distinct events. These lattices form a database of potential paraphrasing techniques or transformations where the areas of high variability represent the event arguments (e.g., the actors, location, etc.). They then cluster sentences written on the same day in different articles to decide which are paraphrases and use the lattice associated with one sentence to generate paraphrases for another sentence in the same cluster – probabilistically, any path through the lattice, with appropriate entity substitution, is considered a paraphrase. Other techniques exist for lexical paraphrase acquisition, which do not rely on parallel corpora (e.g., Glickman and Dagan 2003). The use of patterns extracted a priori by these systems to verify a paraphrase between the learner's answer and the reference answer in a tutoring environment is unlikely to be of much benefit, since they are not broad coverage patterns, but rather are specific to high frequency news topics. However, these algorithms could be modified to perform online paraphrase recognition.

Research in the area of entailment also has much to offer. Lin and Pantel (2001a, 2001b) extract inference rules from text by looking for dependency parse patterns that share common argument fillers according to pointwise mutual information (Church and Hanks 1989). This work stimulated much of the later work in paraphrasing and entailment. The main idea is to find patterns in text that share similar key content words, where these content words fill slot values at each

end of the pattern. The patterns are extracted from paths in a syntactic dependency parse tree and the similarity of patterns is determined by the pointwise mutual information (PMI) computed for the patterns' slots. PMI for a pair of slots indicates whether the sets of words that fill the two slots are more similar than would be expected by chance. This same technique could be applied to determine whether a child's answer to a question is a paraphrase of the reference answer.

Again 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 First Pascal RTE challenge made use of Lin and Pantel's patterns or a variant of their algorithm and still performed quite poorly, barely exceeding chance on any but the easiest task, the comparable documents task, (Braz et al. 2005; Haghighi, Ng and Manning 2005; Herrera, Peñas and Verdejo 2005; Raina et al. 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 binary yes-no judgments as to whether a human reading a text *t* of one or more full sentences would typically consider a second, hypothesis, text *h* (usually one shorter sentence) to most likely be true. Fig. 3 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, the potentially quite long, *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.

Fig. 3. Example text hypothesis pair from the RTE challenge

However, in the third RTE challenge, there was an optional pilot task that begins to address some of these issues. Specifically, they have extended the task by including an Unknown label, where *h* is neither entailed nor contradicted, and have requested justification for decisions. The form that these justifications take is left up to the groups participating, but could conceivably provide some of the information about which specific facets of the hypothesis are entailed, contradicted and unaddressed.

Submitters to the RTE challenge take a variety of approaches including purely 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. 2006), graph matching with semantic roles (e.g., Haghighi, Ng and Manning 2005), logical inference (e.g., Tatu and Moldovan 2007), discourse commitment – assertion,

presupposition, and conversational implicature – strategies (Hickl and Bensley 2007), and numerous other approaches, most based on machine learning. Many of these systems make use of the lexical similarity metrics discussed earlier in this chapter, or lexical relations described more formally in WordNet, or more formally still in most of the logical inference or abductive reasoning systems.

The best performing systems in the RTE challenge have been the Hickl et al. (2006, 2007) approaches. In their 2006 entry, they perform a lexical alignment and then generate four Boolean semantic role features indicating roughly whether the predicates have the same semantic roles and are aligned similarly. Their lexical alignment features include cosine similarity, word co-occurrence statistics, WordNet similarity metrics, NE and POS similarity, and string-based similarity. They consider arguments to probabilistically match if the hypothesis' argument head is lexically aligned with (entailed by) something in the corresponding text's argument. A second set of features includes simpler lexical alignment information based on the longest common substring, the number of unaligned chunks and web-based lexical co-occurrence statistics. A third set of features indicates whether the polarity of the two text fragments is consistent. Their final set of features indicates whether the two text fragments would be grouped together when running a text clustering algorithm on a set of documents retrieved from a query on their keywords. Using a decision tree classifier, Hickl et al. achieved the best results at the second RTE challenge, with an accuracy of 75.4%. Hickl and Bensley (2007) extended this approach at RTE3, extracting the authors' discourse commitments from each text fragment and basing their entailment

strategy on these. Each sentence from the text was elaborated into a potentially very long list of propositions determined to be true based on the original statement's assertions, presuppositions, and conversational implicatures. These commitments were then used in a system based on their RTE2 submission to achieve an accuracy of 80%.

In the following chapters, I build on many of the techniques described in this and preceding chapters, and more importantly to my cause, I develop a more expressive representation framework for recognizing entailment in automated tutoring systems that will lead to more effective dialog and improved learning.

5 Research Overview

Imagine that you are an elementary school science tutor and that rather than having access to the student's full response to your questions, you are simply given the information that their answer was correct or incorrect, a yes or no entailment decision. Assuming the student's answer was not correct, what question do you ask next? What follow up question or action is most likely to lead to better understanding on the part of the child? Clearly, this is a far from ideal scenario, but it is roughly the situation within which many Intelligent Tutoring Systems exist today.

The overarching thesis of this work is that a more detailed assessment of learners' dialog contributions will enable tutoring strategies that will significantly improve learner comprehension. The thesis that this work more directly addresses is that, with the use of fully automated systems, learner contributions can be classified at a level assumed to be appropriate for achieving the above goal of improving learner comprehension and that this assessment can be performed in a domain-independent manner. Such a level of analysis would have to meet the following criteria:

- It must utilize an intermediate level of representation for the reference and learner answers that goes beyond the bag-of-words approach in order to account for the semantic relationships among concepts and that also goes beyond sentence-level analysis in order to provide a more detailed assessment of the learner's understanding.

- The representation must facilitate meaningful assessment of the learner's response at a finer-grained level than a simple correct-incorrect or yes-no entailment decision.
- The representation and assessment must be learnable by an automated system.
- The assessment must not require the handcrafting of domain-specific logic representations, parsers, knowledge-based ontologies, or dialog management rules.

This thesis represents just such a paradigm shift in the assessment of learner responses. In order to achieve this paradigm shift, the work addresses the following primary research questions:

- What type of representation might allow more productive tutoring dialog?
- Can this representation be annotated consistently by human judges?
- Can a machine learning algorithm be trained to generate this annotation automatically?
- Can the algorithm learn to assess learners' answers to questions not seen in the training data or to questions outside the domain on which the algorithm was trained?

This thesis also investigates whether such an algorithm can learn to assess roughly sentence-length answers to science questions from elementary school aged children. This is an area where no known prior work has been successful.

In chapter 6, I address the issue of granularity by designing a new representation scheme capable of providing a fine-grained analysis of student's responses to questions. This representation facilitates the automated tutor's ability to recognize specifically what facets of the reference answer the tutor should focus on during the follow up dialogue and provides much needed detail regarding the student's apparent understanding of those facets. I focus strictly on the student's understanding of the various facets of the reference answer and leave as future work the need to address other issues, such as tangential misconceptions or inaccurate beliefs held by the student.

I employ a supervised machine learning approach in creating the answer assessment component. The process involved in building and using this component is as follows. First, it is necessary to construct the classifier based on gold standard annotated data. In chapters 7, 8 and 9, I describe the corpus I utilize and details of the annotation project. Then I train a classifier to predict the gold standard annotation labels for each facet of the reference answer. This requires that, for each facet, I extract features from the corpus examples that are indicative of the student's understanding of those reference answer facets. These feature vectors are utilized by a machine learning algorithm to build the classifier model. The features and classifier training are detailed in chapter 10. Finally, given a student's response to one of the questions, for each reference answer facet I extract the same set of features used in training and feed these to the classifier which outputs a label categorizing the student's understanding of each facet. This system has not yet been incorporated into an automated tutor, so I evaluate the

system on held out test sets from the original corpus. I describe the associated experiments, present the results, and include a discussion in chapters 11 through 14.

6 Representing Fine-grained Semantics

In order to optimize learning gains in the tutoring environment, there are myriad issues the tutor must understand regarding the semantics of the student's response. Here, I focus on drawing inferences regarding the student's understanding of the low-level concepts and relationships or *facets* of the reference answer. I use the word facet throughout this thesis to generically refer to some part of a text's (or utterance's) meaning. The most common type of answer facet discussed is the semantics associated with a pair of related words and the relation that connects them.

Rather than have a single yes or no entailment decision for the reference answer as a whole, (i.e., does the student understand the reference answer in its entirety or is there some unspecified part of it that we are unsure whether the student understands), I instead break the reference answer down into what I consider to be approximately its lowest level compositional facets. This roughly translates to the set of triples composed of labeled (typed) dependencies in a dependency parse of the reference answer. In a dependency parse, the syntactic structure of a sentence is represented as a set of lexical items connected by binary directed modifier relations called dependencies. The goal of most English dependency parsers is to produce a single projective tree structure for each sentence, where each node represents a word in the sentence, each link represents a functional category relation, often labeled, between a governor (head) and a subordinate (modifier), and each node has a single governor (c.f., Nivre and Kubler 2006). Each dependency can be labeled with a type, (e.g., subject, object,

nmod – noun modifier, vmod – verb modifier, sbar – subordinate or relative clause, det – determiner).

The following illustrates how a simple reference answer (4) is decomposed into the answer facets (4a-d) derived from its dependency parse, with (4a'-d') providing a gloss of each facet's meaning. The dependency parse tree is shown in Fig. 4. As can be seen in 4b and 4c, the dependencies are augmented by thematic roles (e.g., Agent, Theme, Cause, Instrument, etc; c.f., Kipper, Dang and Palmer 2000). The facets also include those semantic role relations that are not derivable from a typical dependency parse tree. For example, in the sentence "As it freezes the water will expand and crack the glass", *water* is not a modifier of *crack* in the dependency tree, but it does play the role of Agent in a semantic parse.

(4) *The long string produces a low pitch.*

(4a) NMod(string, long)

(4b) Agent(produces, string)

(4c) Product(produces, pitch)

(4d) NMod(pitch, low)

(4a') There is a long string.

(4b') The string is producing something.

(4c') A pitch is being produced.

(4d') The pitch is low.

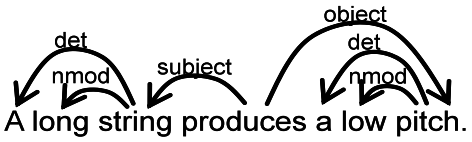

Fig. 4. Dependency parse tree for example reference answer

Breaking the reference answer down into low-level 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? Did they express a related concept that indicates a misconception? Did they leave the facet unaddressed? Can you assume that they understand the facet even though they did not express it, (e.g., it was part of the information given in the question)? It is clear that, in addition to breaking the reference answer into fine-grained facets, it is also necessary to break the annotation labels into finer levels in order to specify more clearly the relationship between the student's answer and the reference answer aspect. There are many other representational issues that the system must be able to handle in order to achieve near optimal tutoring, but these two – breaking the reference answer into fine-grained facets and utilizing more expressive annotation labels – are the emphasis of this thesis.

This chapter provided only an overview of the semantic representation that will be elaborated in the coming chapters. In the next chapter, I discuss the corpus I had annotated according to this representation. Then I describe the

annotation itself, filling in the detail omitted in this chapter, examining some of the issues, and discussing the future work required to more completely represent and assess the learner's understanding of the concepts the tutor is covering. In the chapters following that, I describe the features extracted from this annotated corpus and the training of a machine learning classifier to automatically assess students' responses.

7 Corpus

Because most text comprehension problems take root in elementary school during the early years of learning to read and comprehend text, this work focuses on those critical grades, K-6. I acquired data gathered from 3rd-6th grade students utilizing the Full Option Science System (FOSS), a proven research-based system that has been in use across the country for over a decade (Lawrence Hall of Science 2005). Assessment is a major FOSS research focus, a key component of which is the Assessing Science Knowledge (ASK) project, "designed to define, field test, and validate effective assessment tools and techniques to be used by grade 3–6 classroom teachers to assess, guide, and confirm student learning in science" (Lawrence Hall of Science 2006).

FOSS includes sixteen diverse science teaching and learning modules (see Table 2) and for each module, the FOSS research team designed a set of summative assessment questions with reference answers. These assessments included multiple choice questions, fill in the blank questions, and questions requesting drawings, as well as constructed response questions. I reviewed all of ASK's constructed response questions and selected all of those that were in line with my research goals, which consisted of 287 questions. A representative sample of the questions selected with their reference answers and an example student answer are shown in Table 3.

Grade	Life Science	Physical Science and Technology	Earth and Space Science	Scientific Reasoning and Technology
3-4	HB: Human Body	ME: Magnetism & Electricity	WA: Water	II: Ideas & Inventions
	ST: Structure of Life	PS: Physics of Sound	EM: Earth Materials	MS:Measurement
5-6	FN: Food & Nutrition	LP: Levers & Pulleys	SE: Solar Energy	MD: Models & Designs
	EV:Environments	MX: Mixtures & Solutions	LF: Landforms	VB: Variables

Table 2. FOSS / ASK Learning and Assessment Modules by Area and Grade

These questions had expected responses ranging in length from moderately short verb phrases to several sentences. I eliminated fill in the blank questions and questions that I thought were likely to result in short noun phrase answers regardless of the length of the reference answer, assuming these could generally be successfully assessed by most of today's systems and would not require the approach described in this thesis. Examples of such questions from the Physics of Sound module along with their reference answers and example student responses follow.

HB	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.
ST	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.
ME	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.
PS	Q:	Kate said: "An object has to move to produce sound." Do you agree with her? Why or why not?
	R:	Agree. Vibrations are movements and vibrations produce sound.
	A:	I agree with Kate because if you talk in a tube it produce sound in a long tone. And it vibrations and make sound.
WA	Q:	Anna spilled half of her cup of water on the kitchen floor. The other half was still in the cup. When she came back hours later, all of the water on the floor had evaporated but most of the water in the cup was still there. (Anna knew that no one had wiped up the water on the floor.) Explain to Anna why the water on the floor had all evaporated but most of the water in the cup had not.
	R:	The water on the floor had a much larger surface area than the water in the cup.
	A:	Well Anna, in science, I learned that when water is in a more open are, then water evaporates faster. So, since tile and floor don't have any boundaries or wall covering the outside, the water on the floor evaporated faster, but since the water in the cup has boundaries, the water in the cup didn't evaporate as fast.
EM	Q:	You can tell if a rock contains calcite by putting it into a cold acid (like vinegar). Describe what you would observe if you did the acid test on a rock that contains this substance.
	R:	Many tiny bubbles will rise from the calcite when it comes into contact with cold acid.
	A:	You would observe if it was fizzing because calcite has a strong reaction to vinegar.

Table 3. Sample Questions (Q) from FOSS-ASK with their reference (R) answer and an example student answer (A)

Question: *Besides air, what (if anything) can sound travel through?*

Reference Answer: *Sound can also travel through liquids and solids. (Also other gases.)*

Student Answer: *A screen door*.

Question: *Name a property of the sound of a fire engine's siren*.

Reference Answer: *The sound is very loud. OR The sound changes in pitch.*

Student Answer: *Annoying*.

I also eliminated questions that could not be assessed objectively or that were very open ended. Examples of such constructed response items are:

Question: *Design an investigation to find out a plant's range of tolerance for number of hours of sunlight per day. You can use drawings to help explain your design.*

Question: *Design a way to use carbon printing to find out if two Labrador retrievers have the same paw patterns. Be sure your plan will not be harmful to the dogs.*

Still, there were several moderately open ended questions within the 287 selected. Generally, open ended questions were included if it seemed highly likely that students would address the same points that were included in the reference answer. An example of a question in this category follows.

Question: *What should you do if it appears that an animal is being harmed during an investigation?*

Reference Answer: *Answers will vary. Examples: Be more careful with the animal. Stop the investigation. Change the investigation so it is safer for the animal.*

I generated a corpus from a random sample of the students' handwritten responses to these questions. ASK was pilot tested in a number of schools across the U.S. and in Canada, with each ASK module typically being tested in two to five schools. Therefore, the students whose answers were transcribed represent a reasonably broad spectrum of the population. The only special transcription instructions were to fix spelling errors (since these would be irrelevant in a spoken dialog environment, the target of this work), 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* (since these are not particularly interesting from the research perspective).

In total, approximately 16,000 student responses were transcribed, roughly 100 per question for three test set modules (Environment, Human Body and Water) and 40 per question for the remaining thirteen modules. This resulted in about 144,000 total facet annotations. Three test sets were created by 1) withholding all the data from the three modules discussed above – resulting in a dataset that can be used to test domain-independent algorithms or performance, 2) withholding all answers to a subset of questions from each of the other modules –

resulting in a dataset that can be used to test question-independent algorithms or performance, and 3) withholding approximately 6% of the answers to the remaining questions – resulting in a dataset that can be used to test algorithms intended to handle specific predetermined questions. There are 56 questions and approximately 5,600 student answers in the domain-independent test set, comprising approximately 20% of all of the questions utilized and 36% of the total number of transcribed student responses. There are 22 questions and approximately 880 student answers in the question-independent test set, comprising approximately 8% of all of the questions and 6% of the responses. The third test set spanned the remaining 73% of the questions and included around 500 learner responses or 3.2% of all responses. This resulted in around 45% of the answers being set aside for testing the learning algorithms, with the remainder designated for training and development tuning.

I selected the three domain-independent test set modules because they appeared to be representative of the entire corpus in terms of difficulty and appropriateness for the types of questions that met my research interests. They were also roughly average sized modules in terms of the number of questions they contained. The items included in the question-independent test set were chosen randomly, but with two criteria. First, the items were chosen to include at least one question from each module in the training set and to, otherwise, maintain approximately the same question proportions as the training set (the five smallest modules had only one question, the largest had three, and the remaining seven modules had two questions). Second, I did not include questions whose reference

answers had significant overlap with questions that would remain in the training data. For example, the following questions from the Ideas and Inventions module would not have been selected due to their reference answer similarity.

> Question: *Landra was trying to find out which pen with blue ink was used to write a note in her class. If she used chromatography to find the pen that wrote the note, … explain how she could use the chromatograms from the pens she tested to determine which one wrote the note.*
>
> Reference Answer: *She should compare the pattern of colors on the chromatograms from the pens she tested with a chromatogram from the ink on the note.*
>
> Question: *James had two brown watercolor pens. He wanted to find out if they were made by the same company. He made chromatograms using each of the two pens. How would James use the chromatograms to help him decide if the pens were made by the same company?*
>
> Reference Answer: *James should compare the pattern of the pigments on the chromatograms. If they are similar the pens were probably made by the same company.*

In order to maximize the diversity of language and knowledge represented by the training and test datasets, random selection of students was performed at the question level rather than using the same students' answers for all of the

questions in a given module. However, in total there were only about 200 children that participated in any individual science module assessment, so there is still moderate overlap in the students from one question to another within a given module. On the other hand, each assessment module was given to a different group of children, so there is no overlap in students between modules.

8 Reference Answer Markup

The annotation of student answers consists of two principal steps. First, each reference answer in the corpus, as specified by the ASK research team, was decomposed by hand into its constituent facets. Then each student answer was annotated relative to the facets in the corresponding reference answer to describe whether and how those facets were addressed by the student. Every student answer was double-blind annotated and a third annotator reviewed the others' labels and made the final decision on each facet's label. I describe the details associated with the reference answer markup in this chapter and the student answer annotation details in the next chapter.

8.1 Reference Answer Decomposition and Representation

The ASK assessments included a reference answer for each of their constructed response questions. These reference answers were broken down into low-level facets, roughly extracted from the relations in a syntactic dependency parse (c.f., Nivre and Scholz 2004) and a shallow semantic parse (Gildea and Jurafsky 2002). This decomposition was performed by hand with the assistance of an undergraduate Linguist, who made the first pass over the majority of the reference answers, with me reviewing and modifying the analysis. Since the decomposition is based closely on well established frameworks, dependency parsing and shallow semantic parsing, it was not included in the scope of the experimental research – no formal guidelines were written and the facets were not annotated double blind to calculate inter-annotator agreement.

The Physics of Sound reference answers were distilled into their most critical elements. However, minimal changes were made to the remaining answers, since it would be desirable for the system to be able to handle future reference answers written by educators who do not have detailed knowledge of the assessment system, and in the long-term, to handle questions and reference answers generated automatically by the ITS. The most common transformations were to replace nearly all pronouns with their coreferring nouns and to occasionally drop small parts of sentences that were not relevant to the key concepts. The following is a typical example that illustrates each of these modifications in italics.

> Original Reference Answer: *James should* compare the pattern of the pigments on the chromatograms. (If *they* are similar the pens were probably made by the same company.)
> Modified Reference Answer: Compare the pattern of the pigments on the chromatograms. If *the chromatograms* are similar the pens were probably made by the same company.

The decomposition of the final reference answers began by determining the dependency parse, following the style of the 2006 version of MaltParser (Nivre et al. 2006) – they have since expanded their dependency tags to include all of the functional tags in the Penn Treebank. This dependency parse was then modified in several ways. First, wherever a shallow semantic parse would identify a predicate argument structure, I used thematic role labels (c.f., Kipper,

Dang and Palmer 2000) between the predicate and the argument's headword, rather than the MaltParser dependency tags. This also involved, adding new structural dependencies that a typical dependency parser would not generate, as discussed in chapter 6. In a small number of instances, these labels were also attached to noun modifiers, most notably the Location label. For example, given the reference answer fragment *The water on the floor had a much larger surface area*, one of the facets extracted was Location_on(*water, floor*).

Various linguistic theories take a different stance on what term should be the governor in a number of phrase types, particularly noun phrases. In this regard, the manual parses here varied from the style of MaltParser by raising lexical items to governor status when they contextually carried more significant semantics. For example, the noun phrases *the pattern of pigments* and *the bunch of leaves* typically result in identical dependency parses. However, in Fig. 5, the word *pattern* is considered the governor of *pigments* and thus also modifies *Compare* and governs *chromatograms*; whereas, in Fig. 6, the word *leaves* is treated as the governor of *bunch* because it carries more semantics and thus becomes a modifier of *are* (or the governor of *part* as described in the next paragraph).

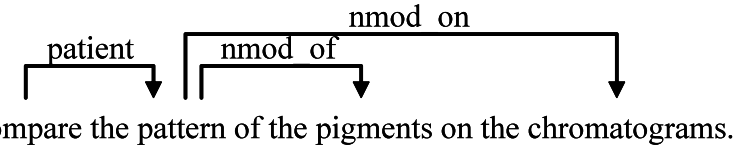

Fig. 5. Typical dependency parse of a NP with an embedded PP

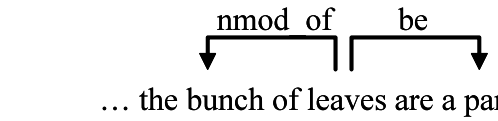

Fig. 6. Nonstandard dependency parse raising core semantic term *leaves* to head

The parses were also modified to incorporate prepositions, copulas, terms of negation, and similar terms into the dependency type labels (c.f., Lin and Pantel 2001). This can be seen in the reference answer fragments in Fig. 5 and Fig. 6, where *of*, *on*, and *are* were incorporated into the relations of the consolidated dependencies, (e.g., normally *pattern of pigments* would be parsed as two dependencies, NMod(*pattern, of*) and PMod(*of, pigments*), but here they are combined into the single dependency NMod_of(*pattern, pigments*)). When auxiliaries did not contribute much to the semantics of the reference answer, they were not included in the facets extracted. Fig. 7 shows the standard MaltParser dependency parse and the revised parse for a reference answer fragment that includes several of the issues discussed in this paragraph. Example 5 illustrates the decomposition of this same answer fragment into its constituent facets along with their glosses.

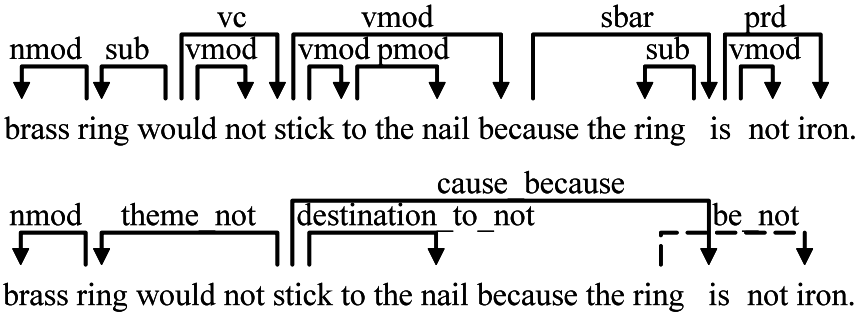

Fig. 7. Typical dependency parse revisions for reference answer facets

(5) The brass ring would not stick to the nail because the ring is not iron.

(5a) NMod(ring, brass)

(5a') The ring is brass.

(5b) Theme_not(stick, ring)

(5b') The ring does not stick.

(5c) Destination_to_not(stick, nail)

(5c') Something does not stick to the nail.

(5d) Be_not(ring, iron)

(5d') The ring is not iron.

(5e) Cause_because(stick, is)

(5e') 5b and 5c are caused by 5d.

I refer to facets that express relations between higher-level propositions as inter-propositional facets. An example of such a facet is (5e) above, connecting the proposition *the brass ring did not stick to the nail* to the proposition *the ring is not iron*. In addition to specifying the headwords of inter-propositional facets (*stick* and *is*, in 5e), I also indicate up to two key facets from each of the propositions that the relation is connecting (b, c, and d in example 5). Reference answer facets that are assumed to be understood by the learner a priori, (e.g., because they are part of the question), are annotated to indicate this. The details of reference answers including their facet definitions are stored in a stand-off markup in an xml file.

There were a total of 2878 reference answer facets, resulting in a mean of 10 facets per question (median of 8 facets). Table 4 shows a high-level break down of the reference answer facets. Facets that were assumed to be understood a priori by students accounted for 33% of all facets and inter-propositional facets accounted for 11%. The experiments in automated annotation of student answers (chapters 12 and 13) focus on the facets that are not assumed to be understood a priori (67% of all facets); of these, 12% are inter-propositional. Fig. 8 charts the frequency of questions that had a specified number of total facets and the frequency that had the specified number of facets not assumed to be understood a priori.

Category of Facets	Frequency	Frequency / Question	% of Total	% (not) assumed
All facets	2878	10.0	100	
Assumed	950	3.3	33	
Not assumed	1928	6.7	67	
Inter-propositional	325	1.1	11	
Simple	2553	8.9	89	
Inter-propositional assumed	100	0.3	3	11
Simple assumed	850	3.0	30	89
Inter-propositional not assumed	225	0.8	8	12
Simple not assumed	1703	5.9	59	88

Table 4. High-level break down of reference answer facets

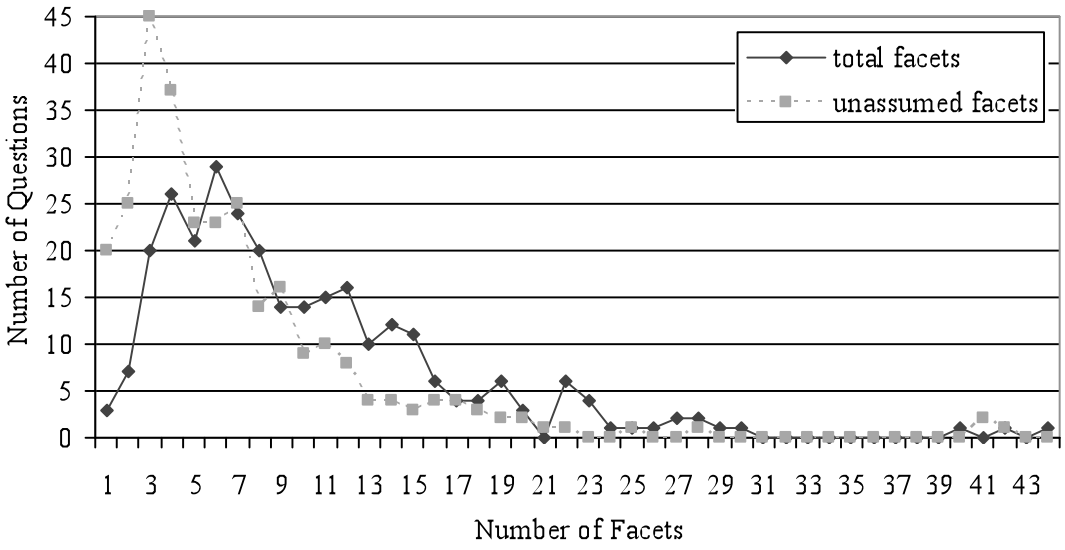

Fig. 8. Frequency of questions with a number of facets

A total of 37 different facet relation types were utilized (see Table 5). The majority, 21, are VerbNet thematic roles. Direction, Manner, and Purpose are PropBank adjunctive argument labels (Palmer, Gildea and Kingsbury 2005). Origin, Quantifier, and Cause-to-Know were added to the preceding thematic roles. Additionally, as indicated above, copulas and similar verbs (e.g., *be, become, do,* and *have*) were themselves considered to be facet relation types connecting their arguments. Finally, anything that did not fit into the above categories retained its dependency parse type: VMod (Verb Modifier), NMod (Noun Modifier), AMod (Adjective or Adverb Modifier), and Root (Root was used when a single word in the answer, typically *yes, no, agree, disagree, A-D,* or a number, stood alone without a significant relation to the remainder of the reference answer; this occurred only 21 times, accounting for fewer than 1% of the reference answer facets). The seven highest frequency relations are NMod,

Theme, Cause, Be, Agent, AMod, and Location, which together account for 75% of the reference answer facet relations (see Fig. 9).

VerbNet Role	Not Asmd	Asmd	Total
Actor	1	0	1
Agent	92	67	159
Attribute	15	3	18
Beneficiary	3	0	3
Cause	159	83	242
Destination	55	17	72
Experiencer	6	1	7
Extent	21	10	31
Instrument	11	5	16
Location	86	44	130
Material	14	5	19
Patient	40	19	59
Predicate	16	3	19
Product	33	13	46
Recipient	9	7	16
Source	10	5	15
Stimulus	13	5	18
Theme	357	155	512
Time	64	20	84
Topic	8	2	10
Value	2	1	3

Other Roles	Not Asmd	Asmd	Total
PropBank Adjs			
Direction	18	5	23
Manner	42	7	49
Purpose	2	2	4
Misc. Types			
Cause-know	15	12	27
Origin	2	0	2
Quantifier	60	26	86
Special Verbs			
Be	144	50	194
Become	7	2	9
Do	1	0	1
Have	23	25	48
Dependencies			
AMod	113	24	137
NMod	447	324	771
Root	21	0	21
VMod	18	8	26

Table 5. Reference answer facet types and frequencies

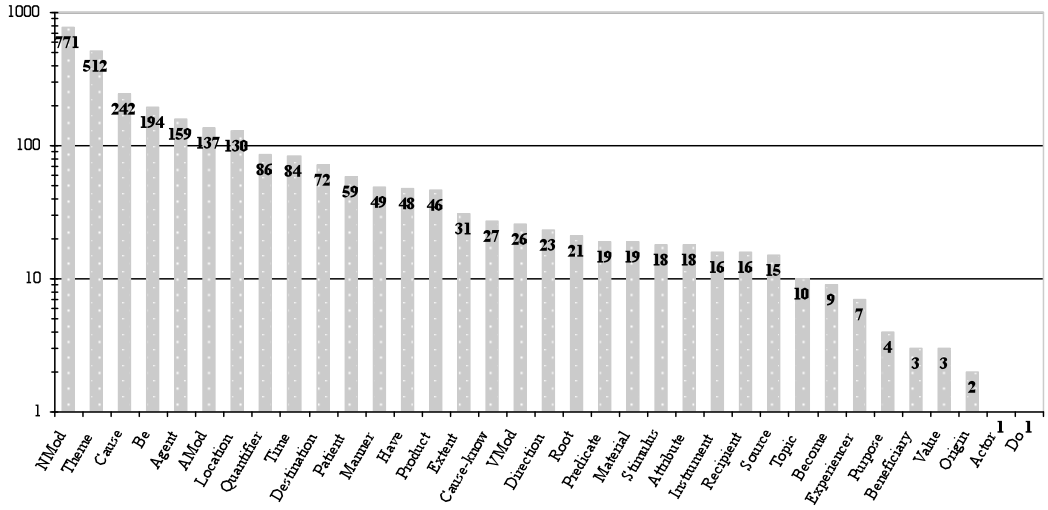

Fig. 9. Logarithmic chart from highest to lowest frequency facet relation types

8.2 Discussion and Future Work on Reference Facets

There are many interesting and open issues in the area of reference answer facet extraction. For example, many adjectives have properties extremely similar to verbs, while others do not. In the first case, I generally extracted facets considering the adjective as a predicate and using the associated thematic roles to potentially distant arguments that would typically not be connected in a dependency parse. For example, in the reference answer fragment *the surface tension is broken*, I treat *broken* as the predicate and extract the facet Patient(*broken, tension*) rather than Be(*tension, broken*). However, the second set of adjectives are treated as simple noun modifiers or as the modifiers in copula relations (e.g., in *oil is less dense*, I extract Be(*oil, dense*)). Other adjectives exist somewhere in the middle of the range (e.g., *longer, shorter, tighter, looser, lower, and saturated* all have verb forms, but they are used infrequently, particularly in the context given by these reference answers). The criteria for when adjectives

should be treated as predicates and when they should be treated as modifiers should be examined more closely. Perhaps the right approach is to use their most common form co-occurring with the other key terms in the sentence (e.g., if *saturated* occurs more often as an adjective than as a verb in the presence of the word *mixture*, it should be treated as an adjective modifying the noun rather than as a predicate governing it).

Further consideration should be given to how to handle conjunctions and disjunctions. Often, it is sufficient to simply extract a separate facet for each conjoined noun, but other times this is inappropriate, (e.g., in *The seed becomes bigger and heavier*, it is okay to treat this as *The seed becomes bigger* and separately, *The seed becomes heavier*; whereas, it is inadequate to treat *salt and water form a solution*, as *salt forms a solution* and, separately, *water forms a solution*).

A more formal analysis should be conducted to decide what characteristics contribute to text adding little value to the reference answer semantics. A common case involves modals; generally modals add very little value to a reference answer and can be omitted from the facets, but at times it is important that they be addressed. For example, in the reference answer *Elena should include a separate shelter for each lizard*, the modal *should* adds relatively little value; whereas, in the answer fragment *If the magnets are weak, the second piece of foam might put too much distance between the two magnets for the interaction to hold the magnets in place*, the modal *might* is more important, since an alternative, incorrect student belief is that the second piece of foam would

definitely add too much distance. Quantifiers result in similar situations; in *The channel for the ships cannot go over any of the shallow areas*, the phrase *any of* can be dropped without loss of significant meaning; whereas, dropping *any* in *Evaporation can occur at any temperature* does result in a loss of important information. If semantically vacuous text was left in the reference answers, rather than revise the first pass reference answer decomposition completely to remove this text as described in the previous section, I often just removed the associated facets. For example, in the reference answer fragment *Insulate each house with a different kind of insulation*, the facet connecting *Insulate* and *insulation* adds no value over the concept of *Insulate* on its own, so I removed this facet from the list.

When a facet is assumed to be understood by a student a priori, it is marked to indicate this. However, currently only the entire facet can be tagged as assumed, while at times it would be optimal to mark one of the terms in the facet as assumed, but not the other. For example, given the question *Describe the function of the seed coat* and the reference response *The seed coat protects the seed until the plant begins to grow*, you can assume that the student will be discussing the seed coat, but you cannot mark the entire facet Agent(*protects, coat*) as assumed, since it includes other key information that you cannot assume to be understood a priori. The ability to tag one term in a facet as assumed should be added to the system in the future.

Additional markup will also be required to ensure appropriate integration of this assessment technology into the ITS. For example, in many cases, there are one or more facets that are not important enough to result in extended dialogue.

There are also cases where the student is only required to address a fraction of the reference answer, (e.g., *Explain how one of the observations helped you to decide whether the stone is a rock or a mineral*).

In the long term, when the ITS generates its own questions and reference answers, the system will have to construct its own reference answer facets. The automatic construction of reference answer facets must deal with all of the issues described in this section and is a significant area of future research. Only with the automatic extraction of reference answer facets can the assessment technology described in this thesis to be considered completely domain-independent.

9 Student Answer Annotation

9.1 Annotation Guidelines

The answer assessment annotation described in this chapter is intended to be a step toward specifying the detailed semantic understanding of a student's answer that is required for an ITS to interact as effectively as possible with a learner. With that goal in mind, annotators were asked to consider and annotate according to what they would want to know about the student's answer if they were the tutor. The key exception here is that they are only annotating a student's answer in terms of whether or not it implies that the student understands each facet of the reference answer. If the student also discusses concepts not addressed in the reference answer, those points are not annotated regardless of their quality or accuracy.

After analyzing much of the Physics of Sound data, I settled on the eight annotation labels noted in Table 6 (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 guidelines and are only very briefly summarized in the remainder of this section.

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

Table 6. Facet Annotation Labels

Example 6 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 (7) along with its final annotation in 7a-c. It is assumed that the student understands that the pitch is higher (facet 6b), 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 6c). Therefore, unless the student explicitly addresses these facets they are labeled *Assumed*.

(6) 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.

(6a) Be(string, tighter), ---

(6b) Be(pitch, higher), Assumed

(6c) Cause(6b, 6a), Assumed

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

(7a) Be(string, tighter), Diff-Arg

(7b) Be(pitch, higher), Expressed

(7c) Cause(7b, 7a), Expressed

Since the student does not contradict the fact that the string is tighter (the string can be both longer and tighter), we do not label this facet as *Contradicted*. If the student's response did not mention anything about either the *string* or *tightness*, we would annotate facet 7a 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 7c, expressing the causal relation between 7a and 7b, is labeled *Expressed*, since the student did express a causal relation between the concepts aligned with 7a and 7b. 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 8, where the student simultaneously expresses the contradictory notions that the string is tighter and that there is less tension.

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

(8a) Be(string, tighter), *Self-Contra*

(8b) Be(pitch, higher), Expressed

(8c) Cause(8b, 8a), Expressed

Example 9 illustrates a case where a student's answer is labeled *Inferred*. 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).

(9) 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.
(9a) Root(root, agree), Expressed
(9b) Be(vibration, movement), *Inferred*
(9c) Agent(produce, vibrations), Expressed
(9d) 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. The first reason is that most systems today cannot hope to detect very many pragmatic inferences, which are the main source of the Inferred and Assumed labels, and

including these in the training data can sometimes confuse learning algorithms resulting in worse performance. 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 to discriminate between these opposing classifications if the classes are distinguished (for algorithms where this is not the case, the classes can easily be combined automatically). Similarly, this allows the possibility of training separate classifiers to handle the different forms of inference. The third reason for separate labels is that it can facilitate system evaluation, including the comparison of various techniques and the effect of individual features – one can assess separately whether a technique or feature had a positive or negative impact on the Inferred facets or on the Expressed facets.

9.2 Annotation Tool

The annotation tool displays the question, reference answer and student answer at the top of the window and, in the main body, lists all of the reference answer facets that students are expected to address (see Fig. 10). Both a formal relational representation and an English-like gloss of the facets are displayed in a table, one facet per row. The annotator's job is to select one of the possible labels (see Table 6) from a drop-down list for each facet to indicate the extent to which the student addressed that facet.

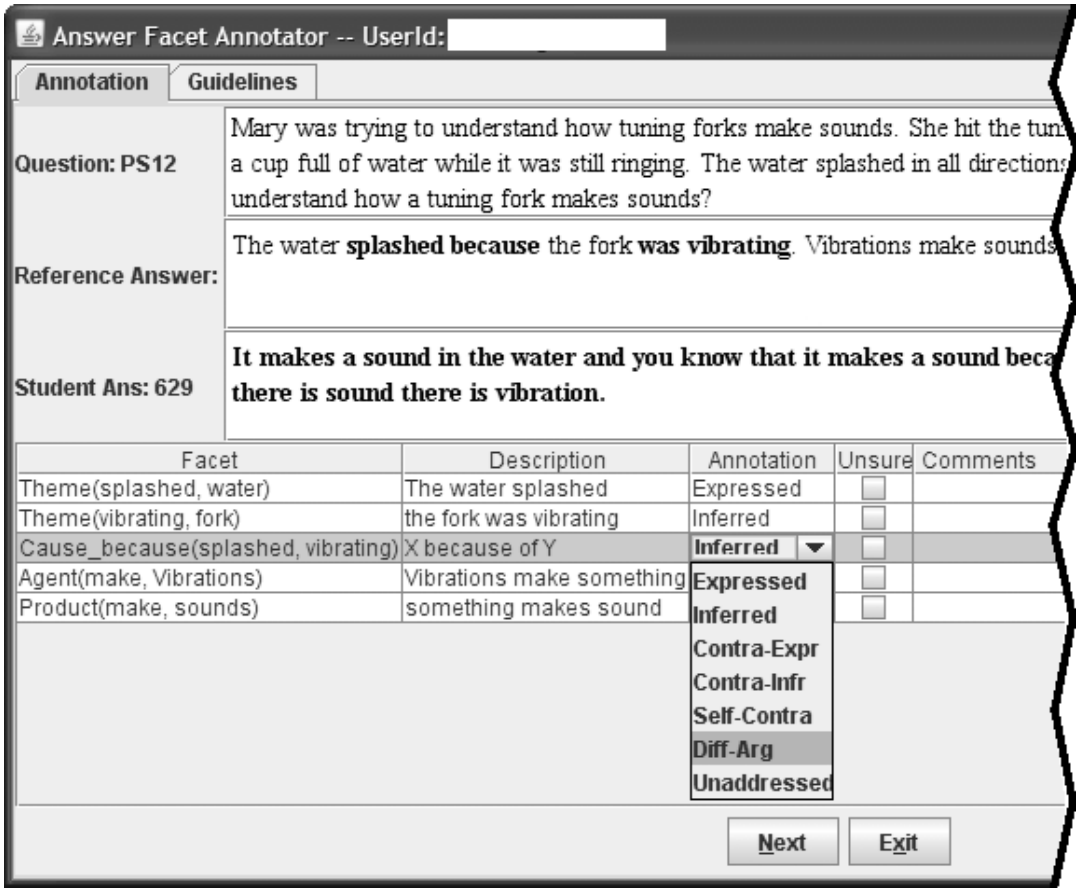

Fig. 10. Partial image of the annotation tool

When the facet is assumed to be understood a priori, the Unaddressed label is not an option in the drop-down list and, since Assumed is essentially the same thing as Inferred a priori, Inferred is also not an option. Originally, the annotation tool automatically defaulted the Assumed label for these facets and annotators were expected to change it as appropriate, but I found that in many cases annotators appeared to forget to review these facets. Similarly, when both annotators had agreed, the tool automatically filled in the labels during adjudication, but again, even in cases where the adjudicator had clearly expressed a different point of view on similar previous student responses, they seemed to

forget to review many of these pre-filled facets. Therefore, I modified the tool to force annotators to physically specify their choices in all cases.

Occasionally it is difficult to know specifically to which part of the reference answer a facet is referring. This is particularly true when a facet might be essentially repeated in a different context within the reference answer or for facets that express relations between other facets or higher-level propositions. To help resolve this ambiguity, when an annotator clicks on a facet, the tool emphasizes the associated parts of the reference answer by setting the font of the key terms to bold (see the facet being annotated in Fig. 10).

The annotated student answers are stored in a stand-off markup in xml files, including an annotated element for each reference answer facet. The annotation tool retrieves one student answer at a time from a web server and returns the annotation for storage in the same xml format.

9.3 Annotators

Annotators were all college students, ranging from first year undergraduates to graduate students and came from a variety of departments including Education, Linguistics, and Computer Science at CU Boulder and Cognitive Science at MIT. In total, seven annotators were involved over the course of the project. Generally, the same annotator performed the entire first, second, or adjudication tagging for all of the questions in a given science module to reduce the learning curve.

9.4 Annotator Training

Training annotators normally took a significant amount of time and the training did not transfer well to new science modules. Initial training was on the Physics of Sound science module. I created three annotator training datasets, the first of which included many of the more challenging-to-annotate student answers, with the second and third sets representing a random distribution of answers. Having read the annotation guidelines, the training consisted of blind annotation of the training set answers with immediate system feedback following the annotation of each answer. This forced the annotators to think through the annotation and rationalize their own thoughts before learning the gold-standard label. Early in the project, the feedback was simply a message indicating which of the facets were labeled inconsistently with the gold standard, under the hope that having to think through what the gold standard might be would result in deeper learning than just presenting the gold standard immediately. The first annotator found this extremely frustrating; so the process was modified such that, after thinking it through and making their own annotation decisions, the gold standard labels were simply displayed in the comments field for the annotator to review. They then had to revise their annotations before proceeding.

The average agreement with the gold standard annotation on the first pass through the initial, challenging dataset was around 80%. After jointly reviewing and discussing the differences, the annotators repeated the same dataset until they felt comfortable with the rationale behind the gold standard annotation. This same process was repeated for the next annotator training dataset. The final

dataset was intended more as a means of estimating expected performance, but still did provide the same training feedback.

Inter-annotator agreement on the final dataset for the first two annotators was 87.7 % and their average agreement with the gold standard was 93.5%. However, inter-annotator agreement on the next module annotated dropped to 77.5%. I believe the primary reason for this was that, while the training annotation review was directed toward the general concepts in the guidelines, in practice, the annotators were also learning question-specific patterns of annotation that could be applied to several very similar student answers. It is also likely that the annotation was somewhat easier for the Physics of Sound module, since I had simplified most of its reference answers. Another factor is that each module required implicit learning of the scientific concepts. For example, annotators needed to (re)learn the scientific usage of terms, such as *effort* and *work*, and the differences between these and their colloquial meanings. There is also a fair amount of scientific terminology that is not commonly used outside of science education, such as cotyledon, isopod, elytra, and chromatogram.

Therefore, hoping to improve inter-annotator agreement on other modules, I decided to create small training sets for each science module. The final process involved selecting six student answers for each question that were representative of a broad spectrum of the more challenging-to-annotate answers. These answers went through the regular annotation process, being blind annotated twice and then adjudicated, but this adjudication, generally performed by me, often included significant detail on why the final label was selected. The adjudicated values

were moved to the first annotation's comments field and these answers were then included as the first set of data to annotate before proceeding with the unannotated answers. In other words, the annotator would see the six training examples, one-by-one, with the first two annotators' tags in their columns and the gold standard annotation visible in the comments field along with a rationale where appropriate. After reviewing each of these, the annotator went on to label the remaining answers for that question.

9.5 Inter-annotator Agreement Results

I evaluate the annotation results under three label groupings: 1) *All-Labels*, where all labels are left separate, 2) *Tutor-Labels* consists of the five labels that will be used by the automated tutor, where Expressed, Inferred and Assumed are combined into a single *Understood* class (i.e., the annotator believes the student understands the facet), Contra-Expr and Contra-Infr are replaced with Contradicted (i.e., the annotator believes the student holds a contradictory view), and the remaining labels are left as is, and 3) *Yes-No*, which is a binary decision, Understood versus all other labels. I calculate inter-annotator agreement based on all 16 of the science modules, totaling 144,716 total facet annotations. I also evaluate this agreement using Cohen's Kappa statistic (Cohen 1960), which contrasts the actual proportion of agreement $P(A)$ with the agreement expected by chance $P(E)$:

$$\kappa = \frac{P(A) - P(E)}{1.0 - P(E)}$$

Tutor-Labels are the labels that will be used by the system, since it is relatively unimportant to differentiate between the types of inference required in determining that the student understands (or has contradicted) a reference answer facet. Agreement on the Tutor-Labels is 86.1%, with a Kappa statistic of 0.728 corresponding with substantial agreement. Agreement is 78.5% on All-Labels and 87.9% on the binary Yes-No decision. These too have Kappa statistics in the range of substantial agreement (see Table 7 for details).

Label Grouping	**ITA %**	**Kappa**
All-Labels	78.5	0.704
Tutor-Labels	86.1	0.728
Yes-No	87.9	0.752

Table 7. Inter-annotator agreement by label groupings, with Kappa statistics

The distribution of the 94,592 facet labels that were adjudicated at the time of writing is shown in Table 8. The most frequent fine-grained label is Unaddressed, at 35.3%, and the majority of the Tutor-Labels indicate the student understood the facet.

Label	Count	% of Total	Count	% of Total
Assumed	24,178	25.6		
Expressed	22,582	23.9	58,750	62.1
Inferred	11,990	12.7		
Contra-Expr	882	0.9	1,377	1.5
Contra-Infr	495	0.5		
Self-Contra	55	0.1		0.1
Diff-Arg	1,037	1.1		1.1
Unaddressed	33,373	35.3		35.3

Table 8. Distribution of annotation labels

An analysis of the inter-annotator confusion matrix indicates that the most probable disagreement is between Inferred and Unaddressed, representing 39% of all the disagreements (the confusion as a percentage of the total disagreements is shown in parentheses in Table 9). The next most likely disagreements are between Expressed and the other Understood labels (Inferred and Assumed), comprising 35% of the disagreements. Confusion between Expressed and Unaddressed is also considerable, representing 10.38% of all the annotator disagreements.

Labels % of Total (% of Errs)	Assumed	Expressed	Inferred	Contra-Expr	Contra-Infr	Self-Contra	Diff-Arg	Unaddressed
Assumed	23.73	3.58 (16.63)	n/a (n/a)	0.08 (0.37)	0.09 (0.43)	0.01 (0.04)	0.35 (1.63)	n/a (n/a)
Expressed		15.72	3.95 (18.34)	0.10 (0.45)	0.06 (0.26)	0.03 (0.14)	0.30 (1.41)	2.23 (10.38)
Inferred			5.28	0.06 (0.28)	0.12 (0.55)	0.01 (0.06)	0.31 (1.44)	8.31 (38.60)
Contra-Expr				0.61	0.15 (0.69)	0.02 (0.10)	0.07 (0.31)	0.29 (1.34)
Contra-Infr					0.11	0.01 (0.03	0.03 (0.16)	0.44 (2.04)
Self-Contra						0.01	0.00 (0.01)	0.01 (0.05)
Diff-Arg							0.41	0.92 (4.25)
Unaddressed								32.61

Table 9. Inter-annotator Confusion Matrix by percent of data and (percent of all disagreements)

Annotator agreement with the gold-standard adjudicated values can be seen as an approximation of annotator performance. The associated confusion matrix along with class-level precision, recall and F-measure are shown in Table 10. The overall annotator accuracy by this estimate is 84.6% on the fine-grained labels shown here and 90.2% on the Tutor-Labels. This is most likely an over-estimate of annotator accuracy, since adjudicators are probably biased toward using the labels chosen by the annotators. This confusion matrix indicates that

annotators are more than twice as likely to errantly choose Unaddressed over Inferred (28.6% of the errors) as to make any other error. The reverse error is the second most likely, with confusions among positive labels following.

Annotator Label: / Gold Label	Assumed	Expressed	Inferred	Contra-Expr	Contra-Infr	Self-Contra	Diff-Arg	Unaddressed	Precision	Recall	F-measure
Assumed	24.37	0.99 (6.40)	n/a (n/a)	0.03 (0.20)	0.02 (0.14)	0.00 (0.02)	0.15 (0.95)	n/a (n/a)	93.1	95.3	0.942
Expressed	1.67 (10.83)	19.75	1.25 (8.11)	0.04 (0.24)	0.02 (0.11)	0.01 (0.07)	0.16 (1.03)	0.98 (6.34)	86.0	82.7	0.843
Inferred	n/a (n/a)	1.63 (10.58)	6.40	0.02 (0.14)	0.03 (0.23)	0.00 (0.03)	0.16 (1.06)	4.42 (28.62)	67.3	50.5	0.577
Contra-Expr	0.02 (0.14)	0.03 (0.23)	0.02 (0.12)	0.67	0.03 (0.21)	0.01 (0.06)	0.02 (0.12)	0.13 (0.81)	71.6	72.1	0.718
Contra-Infr	0.01 (0.09)	0.03 (0.17)	0.04 (0.27)	0.09 (0.55)	0.14	0.00 (0.02)	0.02 (0.15)	0.19 (1.23)	41.0	26.8	0.324
Self-Contra	0.00 (0.01)	0.02 (0.13)	0.01 (0.06)	0.00 (0.02)	0.00 (0.01)	0.02	0.00 (0.00)	0.00 (0.02)	36.9	34.5	0.357
Diff-Arg	0.10 (0.62)	0.07 (0.43)	0.08 (0.54)	0.03 (0.20)	0.01 (0.09)	0.00 (0.02)	0.53	0.28 (1.78)	39.3	48.3	0.434
Unaddressed	n/a (n/a)	0.45 (2.90)	1.70 (11.04)	0.06 (0.38)	0.08 (0.52)	0.00 (0.01)	0.30 (1.97)	32.68	84.5	92.6	0.884

Table 10. Annotator percent agreement with gold-standard and (percent of error)

9.6 Discussion and Future Work on Annotation

Overall, inter-annotator agreement results are reasonable, 86% on the Tutor-Labels, obtaining substantial agreement according to the Kappa statistic. This agreement should be high enough to enable automated classification at a reasonable accuracy. On the other hand, inter-annotator agreement is marginal to

poor on a number of the detailed labels, specifically Self-Contradiction, Different-Argument, Inferred, and Contradiction-Inferred. I examine each of these four labels in the following paragraphs and address why I do not see the lower performance on these four labels as a significant issue. Estimates of annotation errors in this section are drawn from average annotator agreement with the adjudicated gold-standard.

9.6.1 Error Analysis

The Self-Contra class and its estimated errors represent far too little of the data to be a concern, just 0.06% and 0.04% of the data, respectively. However, if annotation errors result in similar system errors and the system interprets these as contradictions (11% of these errors), the dialogue, while not optimal, should not be frustrating or inappropriate, as the student did contradict the facet. If the ITS treats them as Unaddressed or Diff-Arg (11% of the time), it will ask additional questions to clarify the student's understanding, which again is suboptimal, but not a significant problem. If the ITS assumes the student does understand the facet (78% of these errors), the student will miss a potential learning opportunity, but this represents only about 0.04% of all the facets. Furthermore, I believe this is an area that the system might benefit from other labeled training examples, those representing understanding and contradictions. It seems likely that, in the future, the system could outperform humans on this task. It is also worth noting that humans perform extremely poorly on identifying these self-contradictions and yet, human tutoring is extremely effective, implying that a moderate error rate should not be a significant problem.

The Diff-Arg class and its estimated errors account for only 1.1% and 0.6% of the facet annotations respectively. The primary rationale for this class was to attempt to detect misconceptions. There are a variety of views on what misconceptions are; one definition being that a misconception is a belief that is objectively false, is common or typically held by a group of people, and is persistent – has been held for some time or is rooted in an incorrect view of the world that is not trivial to overcome. For example, in a question asking if a brass ring would stick to a magnet, a student indicated that it would because it was made of brass, suggesting a common misconception that all metals can become temporary magnets. Similarly, when asked if an ordinary string would complete a circuit, another student responded *No because a string is not made of iron or steel*, suggesting a common misconception that only iron or steel can conduct electricity. Under this view of misconceptions, the vast majority of facets that fall within the category of Diff-Arg are not misconceptions. However, if the definition is relaxed to simply a belief that is objectively false, then I believe a small majority of the Diff-Arg labels are misconceptions directly related to the reference answer facet. For example, given the partial reference answer *A female* [crayfish] *has an egg pore and longer swimmerets than a male* and the student answer *The female crayfish has an egg pore or belly button, and it has more swimmerets*, the annotation indicates that there is a different argument modifying *swimmerets*, specifically *more* versus *longer*. Furthermore, it does appear that the ITS dialogue could benefit from the recognition that the student addressed a very similar relation and that just a part of the facet was not an appropriate match. For

example, given the reference answer *The seed coat protects the seed until the plant begins to grow* and the student answer *to protect the cotyledon*, the tutor could acknowledge that the seed coat protects something and then focus the dialogue on what exactly it protects.

Therefore, I believe it is worth experimenting with the Diff-Arg tag to see the effect on tutoring quality. In regard to the annotator errors and disagreements associated with Diff-Arg, if it was eliminated, the same annotation errors would exist between its replacement, Unaddressed, and the other classes. The place these errors have the highest likelihood of hurting the dialogue quality is when the student is actually correct. In this case, rather than asking a more vague clarification question, as would be the case with the Unaddressed label, the content of the question would generally convey the system's misunderstanding, which could lead to slightly more frustration on the part of the student. This confusion represents approximately 0.24% of the facets.

Finally and not surprisingly, the agreement on what you can infer a student knows is significantly less than the agreement on whether they have more explicitly expressed that understanding. The Contra-Infr and Inferred classes represent a more substantial 13.2% of the total facet annotations, with their estimated annotation errors accounting for 6.7% of the data. This drops to 4.9% when moving to the Tutor-Labels, (i.e., 1.8% of the confusion is with similar labels, largely Expressed instead of Inferred). Combined the confusion between Inferred and either form of Contradiction and between Contra-Infr and any label indicating the student understood the facet, accounts for only 0.14% of the data.

These confusions will frequently result in poor dialog decisions, but their infrequent occurrence should mitigate the issue. The most common confusions for Inferred and Contra-Infr, representing about 4.8% of the total facet annotations, are with Unaddressed and Diff-Arg, which will each have similar tutoring effects. Many of these errors represent ambiguous student answers. Given the reference answer, *There will be more deposition… Deposition will include larger materials*, the ambiguous student answer *The deposition is bigger…* can lead to disagreements on which facet is Unaddressed and which is Expressed or Inferred between NMod(*deposition, more*) and NMod(*materials, larger*).

My belief with Contra-Infr and Inferred facets is that when an effective probabilistic classifier is trained on these examples, the system's probability estimates will have small margins between the confusable classes when it makes the wrong prediction. In this case, the probability estimates can drive effective dialogue, for example, by gently asking clarification questions when it believes the student did not address the issue or by paraphrasing what it thought the student said during its transition to another question when it infers the student did understand the facet. In cases where the system's classification was wrong, the answer paraphrasing approach would have the effect of clarifying the student's understanding without forcing the larger fraction of students that were correct to repeat or paraphrase themselves a potentially frustrating number of times.

The more important source of errors comes from the classes where annotators had substantial agreement, primarily between Expressed, Contra-Expr,

and Unaddressed (Assumed is similar to Inferred, discussed above). One might think that disagreement between those facets labeled as contradicted and those labeled as understood should be rare, but it turns out that there are numerous reasons why this is not the case. First, there are ambiguous student answers, where two annotators' could have very different interpretations, one suggesting the student understands a facet and the other contradicting it. Second, there are answers that are neither ambiguous nor self contradictions, but yet imply an understanding at some level, and simultaneously a misunderstanding related to the same facet. For example, given the reference answer *Earth materials settle out during deposition* and the student answer *During deposition earth materials would get carried away to a new location*, it is not obvious how to annotate the facet Theme(*settle out, materials*). On the one hand, the student indicates that there is a final destination for the materials in the phrase *to a new location*, which is consistent with this facet and would merit an Inferred label. On the other hand, the student is focusing on the materials being carried away, rather than on them settling, which could imply confusion between deposition and erosion. The third and probably most common source of confusion between these labels results because it is not always clear which facet or set of facets should be labeled as contradicted. Consider the reference answer *Overly Orange takes fewer drops to change the color of indophenol so Overly Orange has a higher concentration of vitamin C* and the student answer *I think that the Overly Orange and the Luscious Lemon do not contain a higher concentration because they clear the indophenol fast*. Here, one annotator said the facet Have(*Overly Orange, concentration*) was

contradicted; whereas, the second annotator labeled it Expressed, but labeled both NMod(*concentration, higher*) and NMod_of(*concentration, C*) as contradicted. The word *not* modifies *contain* in the student's answer, which is consistent with the first annotator's choice, but it seems more logically attached as *not higher*, which is consistent with part of the second annotator's decision. Similarly, given a question about how you know which mineral is harder based on the scratch test, the reference answer fragment *The harder mineral will leave a scratch on the less hard mineral* and the student answer *Whichever one got scratched*, you could either decide that the implied modifiers *soft* and *hard* were reversed and annotate the two associated facets as Contradicted, with the others Expressed and Inferred, or you could consider the action of leaving a scratch to be contradicted and label its two associated facets contradicted, with the above modifiers being Inferred in this context. This alternative labeling would result in four disagreements where one annotator says a facet is Contradicted and the other says it is Understood.

One of the most common sources of annotator error appears to be a bias on the part of annotators toward labeling facets as Unaddressed when, overall, the student did not answer the question correctly. This is corroborated by the fact that there are more than twice as many annotator errors in Table 10 where the annotator chose Unaddressed rather than Expressed or Contra-Expr as there are inverse errors. For example, given the reference answer fragment "… *A steeper slope makes the water flow faster. Faster flow moves more earth materials, increasing erosion.*" and the student answer *I think the slope flatter the water would come down really fast*", both annotators and the adjudicator labeled the

NMod(*Faster, flow*) facet as unaddressed, despite the fact the student clearly states the water would come down really fast. When the student has a better overall understanding, they are more likely to label the facets Expressed, as they did for the above facet given this answer, *Because the more water going at a faster pace because of the slopes will cause more erosion because gravity will help slope*. For the reference answer *The clay particles are lighter or smaller and are therefore carried farther by the water, so the clay particles end up the greatest* ***distance away from the mouth*** *of the river*, annotators were very likely to label the bold face facet as Expressed if the student indicated a large distance and Unaddressed if they indicated a short distance. Similarly, if the student was correct and mentioned *clay*, annotators were inclined to give them credit for *clay particles*; whereas, if the student mentioned *clay*, but there were significant faults in their answer, the annotators labeled *clay particles* as Unaddressed.

Another common source of annotation errors is that taggers were generally very reluctant to draw inferences about the student's understanding unless it was stated fairly explicitly. This too is supported by Table 10 in that there are roughly 2.5 times as many errors where the annotator chose Unaddressed rather than Inferred or Contra-Infr as there are of the inverse errors. For example, given the reference answer, *Colored all the places on the map where the paperclip only went in a short distance. These places are the shallow parts of the harbor. The channel for the ships cannot go over any of the shallow areas.* and the student answer *When to paperclip went in all the way I drew one notch and when it did not go in. I went back and erased and went another way until I reached the dock*,

annotators labeled some facets Unaddressed even though the student seems to understand the principles. However, it is also the case that seemingly trivial inferences are often missed unless the key words are present. For example, given the reference answer fragment ... *The earth materials form a delta when the materials are dropped off or deposited when the water stops flowing at the mouth of a river* and the student answer *A delta is formed when water is deposited*, both annotators and the adjudicator labeled the facet Theme(*deposited, materials*) as Diff-Arg, as they should, but all three also labeled the redundant facet Theme(*dropped off*, *materials*) as Unaddressed.

Annotators also seem to forget the context of the question and often give the student too much credit for simply repeating stated information. For example, given the question, *Kate said: 'An object has to move to produce sound.' Do you agree with her? Why or why not?* and the reference answer *Agree. Vibrations are movements and vibrations produce sound*, student answers such as *I do agree with her because a object has to move to make a sound* often resulted in significant credit despite their lack of any information from outside the question, other than the agreement.

Inter-annotator agreement on the fine-grained labels varied 2.4% among the annotators. It ranged only 0.9% on the Tutor-labels, with the exception of one stand-out annotator, whose agreement was 2.5% higher than the lowest agreement. Investigating the effect of the number of facets on the fine-grained ITA (see Fig. 11), there is no clear trend in the agreement. I anticipated that agreement would fall off when annotators were tasked with passing judgments on

a much larger set of facets, but the graph does not indicate this. However, students tend to address a significantly smaller percentage of the facets in longer reference answers, which presumably makes it easier for annotators to agree on many of the Unaddressed facets. This would imply that the agreement on other labels probably does drop. The more surprising finding is that agreement is much lower for the facets that occur alone in a reference answer. There were three such reference answers: *the bulb will light, the motor will run*, and *moves back-and-forth*.

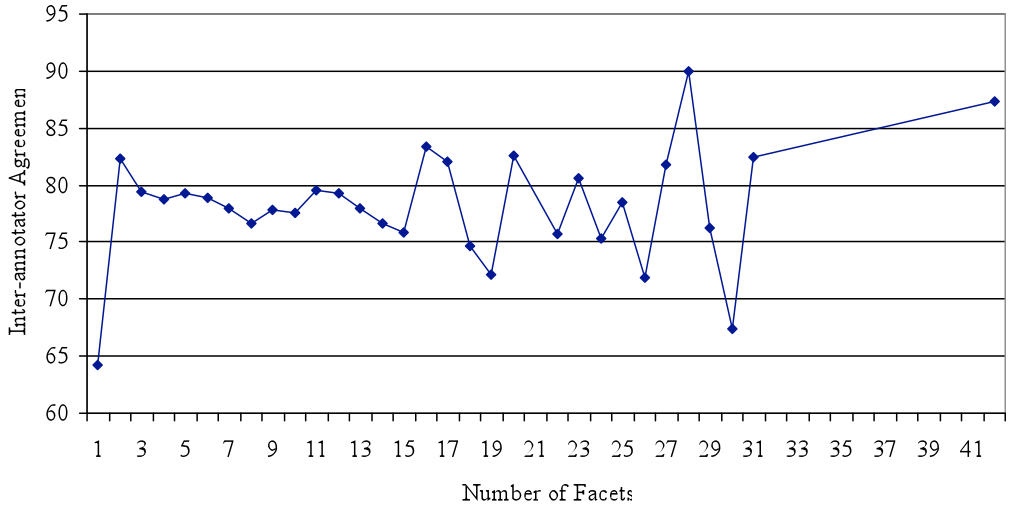

Fig. 11. Inter-annotator agreement by number of facets in the reference answer

Fine-grained agreement varied substantially based on the facet's thematic role, as can be seen in Fig. 12, ranging 28.5% from Value with an agreement of 64.2% (Actor, at only 36.8% agreement, had only a single reference answer facet) to Purpose with an agreement of 92.7%. Dropping the four extremes at each end, the range reduces to 8.7%. Similarly, average agreement varied widely depending on the science module (Fig. 13), with a high of 87.7% for the Physics of Sound

module on which annotators were trained extensively, to 85.1% for Structures of Life, the second highest agreement, down to 71.3% on the module with the worst agreement, Food and Nutrition (Magnetism and Electricity had the lowest agreement on the Tutor-labels, at 81%).

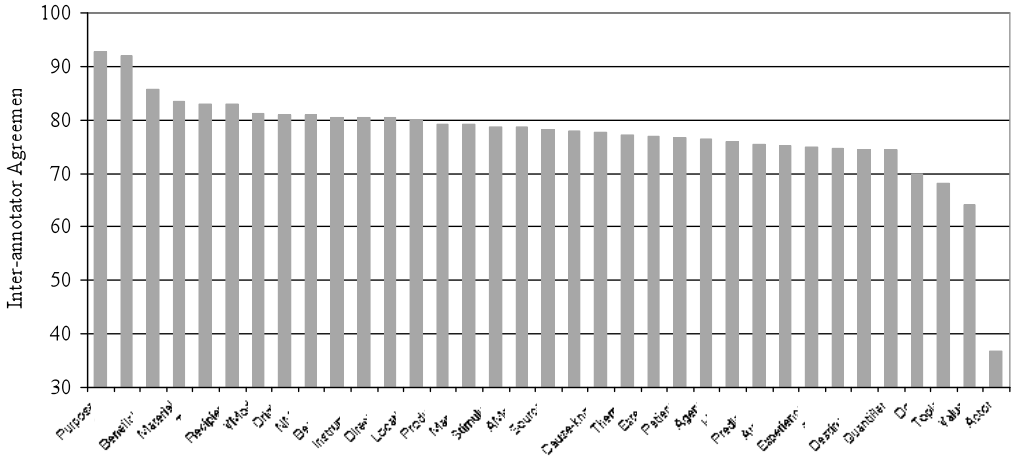

Fig. 12. Inter-annotator agreement by the facet type label

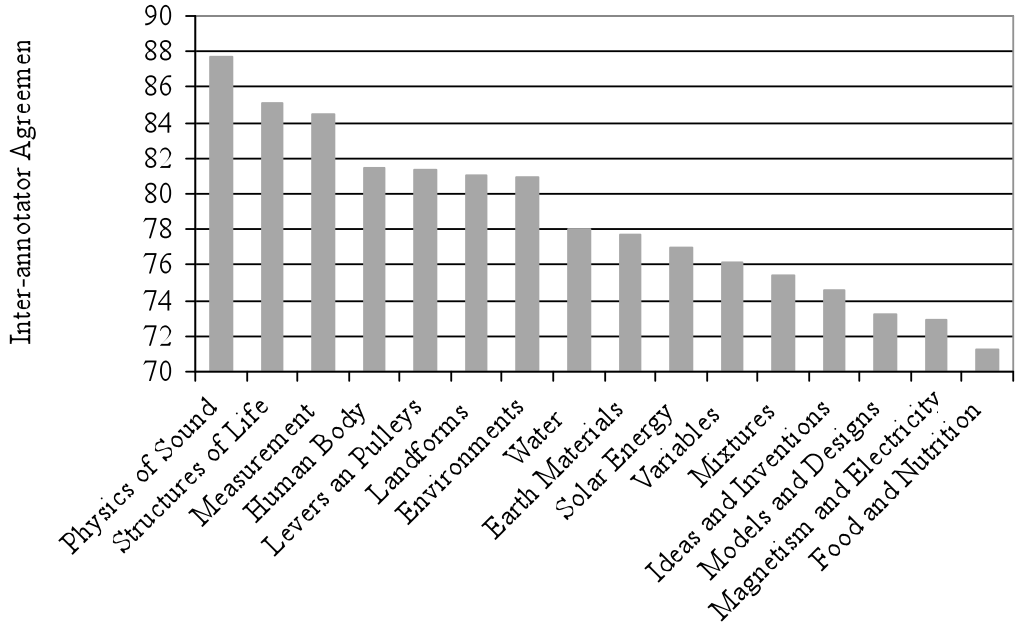

Fig. 13. Inter-annotator agreement by science module

9.6.2 Resolving High Priority Annotation Issues

Future work should include a more thorough analysis of the disagreements between facets labeled contradicted versus understood. It might be possible to find patterns that would result in more consistent annotation and thus, more importantly, more effective tutoring strategies. There are few enough of these disagreements, around 729, that it would not be too arduous a task to modify them if a more consistent strategy emerges. It is tempting to write these errors off as too low a frequency to matter, but these are perhaps some of the most important issues to detect and address in the tutoring session. Another means of addressing these issues might be to supplement the current fine-grained annotation with a secondary annotation that addresses student understanding at a higher propositional level.

In fact, the first version of the annotation guidelines involved annotating the lower-level facets based on their more literal expression and then providing an appropriate overriding annotation at the propositional level where the combination of lower-level annotations was inconsistent with the overall interpretation of the answer. For example, given the reference answer *A longer string produces a lower pitch* and the student answer *A shorter string produces a higher pitch*, the facets NMod(string, longer) and NMod(pitch, lower) would be labeled contradicted, but a higher proposition-level annotation would indicate that in fact the student understood the concept. However, this approach was abandoned since it proved to be too challenging of an annotation task, with the first annotator never making use of the option.

9.6.3 Beyond the Reference Answer

Long term, other aspects of the students' understanding that do not directly relate to the reference answer should also be annotated. Consider example (7), in addition to the issues already annotated, the student contradicts a law of physics that they have surely encountered elsewhere in the text, specifically that longer strings produce lower, not higher, pitches. Under the current annotation scheme this is not annotated, since it does not pertain directly to the reference answer, which has to do with the effect of string tension rather than length. Another example of this issue is seen in a question asking what will happen if a switch is flipped in one direction given a circuit with a switch that can either turn on a motor or a light, but not both. The reference answer is *The bulb will light* and the student answer is *I will start the motor and light*. Here the student expresses the facets of the reference answer, but clearly does not understand the circuit concepts being tested.

In other annotation plans, it would be very useful for training learning algorithms if there is an indication of which student answer facets play a role in each annotation decision. However, I believe this alignment can be done through a combination of unsupervised, semi-supervised and active learning, avoiding the need for further extensive human annotation.

9.7 Annotation Conclusions

The goal of this fine-grained classification is to enable more effective tutoring dialog management. The additional labels facilitate understanding the

type of mismatch between the reference answer and the student's answer. Breaking the reference answer down into low-level facets enables the tutor to provide feedback relevant specifically to the appropriate facet of the reference answer. For example, given the reference answer *In the morning the Sun is in the east so shadows point west. In the afternoon the Sun is in the west so shadows point east.* and the student answer *This happens because in the morning the sun rises north that is why the bus points in the street. The bus points the opposite direction, because the sun sets west.*, the feedback to the dialogue system would indicate that the student contradicted the facet emphasizing the direction of the sunrise. It could then explore this specific issue, rather than asking another general question.

This annotated corpus will benefit researchers in other domains as well. In the question answering domain, this facet-based classification would allow systems to accumulate entailing evidence from a variety of corroborating sources and incorporate answer details that might not be found in any single sentence. In other applications outside of the tutoring domain, this fine-grained classification can also facilitate more directed user feedback. For example, both the additional classifications and the break down of facets can be used to justify entailment system decisions.

The corpus described in this chapter, which should be released in the first quarter of 2008, represents a substantial contribution to the automated tutoring and entailment communities, including 144,716 facet entailment annotations. By contrast, three years of RTE challenge data comprise fewer than 4600 entailment

annotations. There is no other publicly available corpus of elementary school children's answers to any form of education questions. More importantly, this is the only corpus, with learners of any age, which provides entailment information at the fine-grained level described in this thesis. This will enable application development that was not practical previously.

In the remainder of this thesis, I detail how I use the fine-grained annotations described in this and preceding chapters to train a classifier to automatically assess students' understanding of reference answers.

10 Assessment Technology

In this chapter, I describe the initial strategies for assessing a student's understanding of individual reference answer facets. The following chapters present experimental results and discuss plans for future work. A high-level description of the system classification procedure is as follows. Start with the hand-generated reference answer facets described in chapter 8, which are similar to typed dependency triples. Generate automatic parses for the reference answers and the student answers. Then for each student answer, generate a training (or test) example for each facet of the associated reference answer. These examples are comprised of features extracted from the reference and student answers, their dependency parses, and the relation between these. Finally, train a machine learning classifier on the training data and use it to classify the unseen examples in the test sets according to the labels described previously in Table 6.

10.1 Preprocessing and Representation

Many of the machine learning features described here are based on document co-occurrence counts. Rather than use the web as my corpus (as did Turney (2001) and Glickman, Dagan and Koppel (2005), who generate analogous similarity statistics), I use three publicly available corpora totaling 7.4M articles and 2.6B indexed terms.

English Gigaword: English Gigaword (Graff 2003) is newspaper text from five sources ranging from 1995-2004. It consists of about 5.7M news articles and 2.1B words on a wide variety of subjects. This resulted in documents

with an average of around 375 indexed tokens. This corpus comprises 77% of the total documents and 83% of the total indexed words.

Reuters Corpus Volume 1: The Reuters corpus (Lewis et al. 2004) consists of one year of Reuters newswire from 1996-1997. It provided 0.8M articles and 0.17B indexed words, averaging 213 words per article.

TIPSTER: The three volume TIPSTER corpus includes documents from a variety of sources, including newspaper text, and ranges from 1987-1992. It provided 0.9M articles and 0.26B indexed words, averaging 291 words per article.

These corpora are almost exclusively drawn from the news domain, making them less than ideal for assessing student's answers to science questions. This will be addressed in the future by indexing more relevant information drawn from the web. However, the use of these corpora will provide support for the hypothesis that domain-independent assessment is feasible.

The above corpora were indexed and searched using Lucene, a publicly available Information Retrieval tool.[5] Two indices were created, the first using Lucene's StandardAnalyzer and the second adding their PorterStemFilter which replaces the surface form of words with their lexical stem (e.g., vibrate, vibrates, vibrated, and vibrating are all mapped to the same stem). Each index excludes only three words, {*a, an, the*}. However, when referring to content words in the feature descriptions that follow, Lucene's standard stop-word list is utilized, with the exception of removing the words *no* and *not* from their list.

---

[5] http://lucene.apache.org/

There are several natural language processing steps that must be performed before I ultimately extract features to train the machine learning classifiers. First, the answers must be processed by a classifier that performs sentence segmentation. Then the text of sentences is tokenized, breaking it into the discrete linguistic units (e.g., words, numbers, punctuation) required by downstream algorithms. Next, the tokenized sentences are processed by a part-of-speech (POS) tagger. Finally, the POS-tagged sentences are provided as features to generate dependency parses of the reference answers and student answers using MaltParser (Nivre et al. 2007). These parses are then automatically modified to increase the semantic content of the dependencies.

First, auxiliary verbs and their modifiers are reattached to the associated regular verbs. In the example in Fig. 14, this involves reattaching the modal *would*, its subject *ring*, and the verb modifier *not* to the main verb *stick*, making it the new root of the dependency tree. Prepositions are incorporated into the dependency relation labels following Lin and Pantel (2001b). In the example, this results in *nail* being reconnected to *stick* and setting its relation type to VMod_to, the conjunction of the preposition's relation type, VMod, and the preposition itself. Likewise, the copula *is* in the subordinate clause is also reattached to *stick* and is given the relation type VMod_because. Then copulas are incorporated into the dependency relations; the non-subject modifiers of the copulas are reattached to the subject and the relation type between the predicate and subject is updated to incorporate the copula. In the example, this means *iron* and the second instance of *not* are both reattached to the second instance of *ring*. The relation type

connecting *iron* to *ring* is set to Copula_be_prd, reflecting its original predicate role and the incorporation of the copula *is*. In a similar modification, negation terms are appended onto the relevant dependency relations. In the example, both instances of *not* are appended to the dependency relation types of each of their siblings.

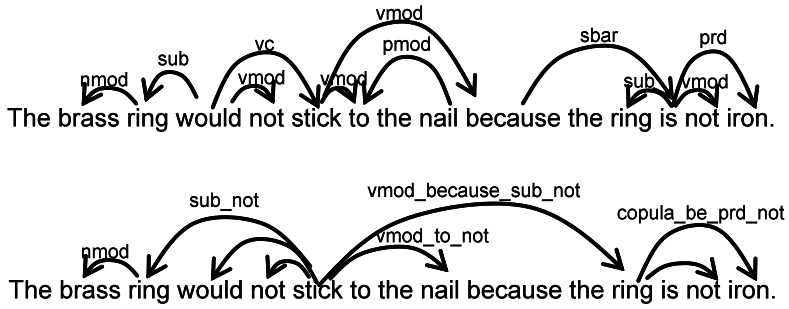

Fig. 14. Example dependency parse tree transformation from top to bottom

These modifications increase the likelihood that terms carrying significant semantic content are joined by dependencies that will be the focus of later feature extraction. For example, in Fig. 14, rather than maintaining the semantically empty dependencies Sub(*would*, *ring*), VMod(*would, not*) and VC(*would, stick*), the transformation results in the much more meaningful dependency Sub_not(*stick*, *ring*). Similarly, the set of dependencies Sub(*is, ring*), VMod(*is, not*), and Prd(*is, iron*), each of which have little meaning in isolation, are reduced to the single dependency Copula_be_prd_not(*ring, iron*), which carries far more significance. The importance of these transformations will become clearer in later sections where features are extracted from these dependency trees.

Lexical Features
Gov/Mod_MLE: The lexical entailment probabilities for the reference answer facet's governor and modifier following (Glickman, Dagan and Koppel 2005; see also, Turney 2001)
Gov/Mod_Match: True if the governor's (modifier's) 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 facet in question represents a relation between propositions
Syntactic Features
Gov/Mod_POS: The part of speech (POS) tags for the facet's governor and modifier
Facet/AlignedDep_Reltn: The dependency or role type labels of the facet and the aligned learner answer dependency
Dep_Path_Edit_Dist: The edit distance between the dependency path connecting the facet's governor and modifier and the path connecting the aligned terms in the learner answer
Other Features
Consistent_Negation: True if the facet has a negation and the aligned learner dependency path has a single negation or if neither have a negation
RA_CW_Cnt: The number of content words in the reference answer, motivated by the fact that longer answers were more likely to result in spurious alignments

Table 11. Machine learning feature descriptions

10.2 Machine Learning Features

I investigated a variety of linguistic features and settled on the features summarized in Table 11, informed by training set cross validation results from a Decision Table (Kohavi 1995). Many of the features dropped provided significant value over the simple lexical baseline, but did not improve on the more

informative features described here. However, I only describe in detail these final features, which provide a solid initial system performance. The features assess lexical similarity via part of speech (POS) tags, lexical stem matches, and lexical entailment probabilities following (Turney 2001; Glickman, Dagan and Koppel 2005). They include dependency parse information such as relevant dependency relation types and path edit distances. Other features include information about polarity among other things. In the rest of this section, I describe some aspects of these features in more detail; this description can safely be skipped on a first reading.

Gov/Mod_MLE: The reference answer facets are comprised primarily of the governing term and its modifier derived from semantic and syntactic dependencies, as discussed in chapter 8. For example, given the reference answer *A paperclip is harder than a penny*, the facets extracted are Be(*paperclip, harder*) and AMod_than(*harder, penny*). The core features assess the likelihood that these two terms, the governor and modifier – *harder* and *penny* in the preceding facet, are discussed in the learner's answer. The features come primarily from the lexical entailment calculations in (Glickman, Dagan and Koppel 2005). Here a lexical entailment probability is derived from maximum likelihood estimates (MLEs) based on corpus co-occurrence statistics. For a single content word *w* from the reference answer *r*, their method estimates the probability of lexical entailment as:

$$P(Tr_w = 1 \mid l) \approx \max_{v \in l} P(Tr_w = 1 \mid v) \approx \max_{v \in l} \frac{n_{w,v}}{n_v}$$

where Tr_w is the truth value of w, l is the learner answer, v represents a word in l, n_v is the number of documents in a predefined corpus in which v occurs (this corpus included GigaWord, Reuters, and Tipster in this thesis), $n_{w,v}$ is the number of documents in which w and v co-occur, and the truth value or entailment of w is assumed to be determined primarily by the single aligned word from l that maximizes this estimate. (Glickman, Dagan and Koppel apply these methods in the RTE challenge, not to learner answer assessment; the reference answer takes the role of the RTE *hypothesis* and the learner answer takes the place of the RTE *text*.) This value is the maximum likelihood estimate that the reference answer term w will occur in a random document given that the document contains the learner answer term v. Rather than consider this a lexical entailment probability, I simply look at it as evidence for the relation of the two words and let the machine learning algorithm decide its role and significance.

Turney applied this metric, calling it PMI-IR, to solve the Test of English as a Foreign Language (TOEFL) synonym task (2001). PMI-IR outperformed Latent Semantic Analysis in Turney's experiments, which in turn achieved results that were statistically as good as the performance of college-aged non-native English speakers on the questions evaluated (Landauer and Dumais 1997).

Given the reference answer *A paperclip is harder than a penny* and the student answer *I know because a penny is softer than a paperclip*, the content words of the reference answer are *paperclip, harder, than,* and *penny*, which align with *paperclip, softer, than,* and *penny* respectively from the learner answer. The exact matching terms are all given co-occurrence probabilities of 1.0. The word

co-occurrence probability of the remaining alignment, *harder* to *softer*, is $n_{harder,softer} / n_{softer} = 457 / 11978 = 0.038$. Since these estimates are subject to significant variance depending on, among other things, n_v – the number of documents in which the potentially entailing word occurs, I expose this count to the classifier so that it can learn how much to trust the estimates. Including this and similar features that are indicative of the validity of co-occurrence statistics appeared to help in the RTE challenge (Nielsen, Ward and Martin 2006), but in evaluating the student answers, did not seem to provide additional value according to feature selection. Two sets of these co-occurrence lexical similarity features were generated for each of the governing term and the modifier. The first set was based on the surface morphological form of the words and the second set was based on their stems.

Glickman et al. take the product of the lexical entailment probability over all content words to generate a single entailment probability for the hypothesis:

$$P(Tr_r = 1 \mid l) = \prod_{w \in r} \max_{v \in l} \frac{n_{w,v}}{n_v}$$

One weakness of this product function is that longer hypotheses (reference answers) result in lower entailment probabilities. Therefore, in the RTE challenge, in addition to the product I also included features for the average and the geometric mean of the probabilities. These features were included here as well, under the assumption that the less the learner answer addresses overall, the less likely it is that they understand any given facet of the reference answer.

These features all seemed to add value in the RTE challenge task, but did not appear to be helpful in assessing student answers.

Gov/Mod_Match: Boolean features were included indicating whether there was an exact match for the governor and modifier in the learner answer. These are essentially redundant with the Gov/Mod_MLE features, since they are true precisely when the MLE features are 1.0 and false otherwise. The rationale for including such features is that it might simplify the machine learning algorithm's task. This appears to be the case as some of these features were chosen in the feature selection.

Subordinate_MLEs: When the reference answer facet represents a relation between higher-level propositions, I include estimates of the dependency, governor, and modifier co-occurrence MLEs for up to two facets associated with the propositions headed by each of the governor and modifier of the current facet. For example, consider the reference answer fragment *The string* is_1 *tighter, so the pitch* is_2 *higher* and its causal facet Cause_so(is_2, is_1). Here, the causal facet relates the proposition *the pitch is higher* to its cause, *the string is tighter*, which each consist of a single facet, Be(*pitch, higher*) and Be(*string, tighter*) respectively. The feature set for the causal facet includes dependency co-occurrence MLEs for each of these facets, plus lexical co-occurrence MLEs for the governor and modifier in each of these facets. The dependency co-occurrence MLEs are estimated just as with the lexical co-occurrence MLEs, but replacing the simple occurrence of a word with the occurrence of both terms in the facet (or the aligned learner answer dependency) in the same sequence as they occur in the

reference answer and within the same number of words (see Nielsen, Ward and Martin 2006 for details).

Gov/Mod_POS: The part of speech of both the facet's governor and its modifier are included in the feature set.

Facet/AlignedDep_Reltn: The facet relation label and its aligned dependency type are also included as features. Reference answer facets are aligned to dependencies from the learner answer using the dependency co-occurrence MLEs described previously for Subordinate_MLEs.

Dep_Path_Edit_Dist: For each term in the reference answer facet the *N* best alignments to the learner answer are found based on the lexical co-occurrence MLEs (*N*=5 in the later experiments). Then for each learner answer word aligned to the facet's modifier, I find the path through the dependency tree to each word aligned to the facet's governor. Likewise, the path through the reference answer dependency parse that connects the facet's modifier and governor is computed. This path is not necessarily a single step due to parser errors and the construction of facets that do not represent typical syntactic dependencies.

Edit distances are calculated between the reference answer path and each of the aligned learner paths. The paths include the dependency relations with their attached prepositions, negations, etc, the direction of each dependency, and the POS tags of the terms along the path. The calculation applies heuristics to judge the similarity of each part of the path (e.g., dropping a subject had a much higher cost than dropping an adjective). The value of the Dep_Path_Edit_Dist feature is the lowest magnitude edit distance.

Consistent_Negation: This is a Boolean feature, which is true if the facet included a negation and the aligned learner dependency path included a single negation. It is also true when neither the path nor the facet had any negations. Otherwise, it is false.

RA_CW_Cnt: Since longer reference answers are more likely to generate spurious alignments and less likely to be addressed in their entirety, the number of content words in the reference answer is included as a feature for the learning algorithm.

10.3 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 within the same domain or topic areas seen in the training data. (Though the specific questions selected were random within a module, the number of questions selected from each module was proportional to the original distribution of questions.) The third test set, *Unseen Answers*, was created by randomly assigning all of the facets from approximately 6% of the remaining learner answers to a test set, with the remainder comprising the training set.

All of the data in the Unseen Modules test set has been adjudicated; whereas, about half of the remaining data (training data, Unseen Questions and Unseen Answers) has not been adjudicated. I used the most recent annotation of the unadjudicated data for the experiments presented here. Today's automated tutoring systems know in advance the questions, reference answers, and what information a student should be *assumed* to understand a priori whether based on the question context or because it is trivial background knowledge. This is equivalent to the Unseen Answers test set; so, in that scenario, it clearly does not make sense to include the facets tagged as Assumed in the test set for the classifier. The long term goal is for the ITS to generate its own questions, in which case it *will* be useful to classify facets that should be assumed to be understood a priori. This has been left for future work, so all Assumed facets were withheld from the data sets in the present experiments. This selection resulted in a total of 54967 training examples, 30514 examples in the Unseen Modules test set, 6699 in the Unseen Questions test set, and 3159 examples in the Unseen Answers test set.

I evaluated several machine learning algorithms and C4.5 (Quinlan 1993) or Random Forests (Breiman 2001) achieved the best results in cross validation on the training data. Therefore, they were used to obtain results for this new task of automatically annotating low-level reference answer facets with fine-grained classifications.

10.4 Evaluation Metrics

In addition to accuracy (the percent of facets classified the same as the gold-standard human annotation), in some experiments I present the confidence weighted score (CWS). CWS provides an indication of the quality of the classifier's confidence judgments. Loosely speaking, classifiers that are correct more often on their higher confidence judgments score better on CWS. The chance value for CWS is equal to the classification accuracy. It is computed after ranking examples according to the classifier's confidence, from most to least confident. Then the CWS is calculated as the average over the system's precision values up to each point in the ranked list:

$$CWS = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{j=1}^{i} \delta\left(\hat{y}_j = y_j\right)}{i}$$

where n is the number of examples (answer facet annotations) in the test set, i ranges over the examples sorted by decreasing confidence in the classification, and $\delta(z)$ is an indicator function (1 if z is true and 0 otherwise), so the embedded fraction is the precision through the i^{th} example.

11 RTE Experiments

11.1 Experimental Design

In prior work, a similar technique was applied to the PASCAL Recognizing Textual Entailment challenge (Nielsen, Ward and Martin 2006). The RTE datasets are composed of text-hypothesis entailment pairs (see Fig. 3 above) derived from several different task domains. The goal is to determine whether the hypothesis is entailed by the text. In the first RTE challenge, the task domains included Information Extraction (IE), Information Retrieval (IR), Question Answering (QA), Machine Translation (MT), Paraphrasing (PP), Reading Comprehension (RC), and Comparable Documents (CD – similar to Multi-document Summarization). In the second RTE challenge, the task domains included IE, IR, QA, and SUM (Multi-document Summarization). Each dataset is balanced such that 50% of the pairs are entailed and 50% are not entailed. The RTE1 training set is approximately 570 examples and the test set is 800 examples. The RTE2 training and test sets are each comprised of 800 examples, 200 for each task type.

In these experiments, the dependency and word co-occurrence statistics were extracted across all content words and combined in product and average features. Similar features were extracted to expose the coverage of the dependency subtrees rooted at each verb, subject, object, and other critical arguments (see (Nielsen, Ward and Martin 2006) for complete details of these and other features).

A mixture of experts consisting of machine learning classifiers from the Weka package (Witten and Frank 2000) was trained with the final classification by a simple majority vote. A confidence estimate was computed based first on the number of votes and, in the case of ties, by averaging the probability estimates output by the classifiers, which were normalized to be consistent with the classifiers' accuracy on training set cross-validation. For the RTE2 dataset, one classifier was trained for the multi-document summarization, SUM, subset of the data utilizing both the RTE2 SUM training data and the RTE1 CD data and a second classifier was trained for the IE-IR-QA (NonSUM) portion of the data utilizing only the associated RTE2 training data, since this resulted in better performance during cross-validation on the training sets. Similarly, for the RTE1 dataset, two classifiers were trained. A CD classifier was trained strictly on the RTE1 CD training set and a NonCD classifier was trained on the remainder of the RTE1 training set.

11.2 Results

Table 12 shows the results on the RTE1 and RTE2 test data. On the RTE2 dataset, in addition to the simple accuracy (percent correct), the average precision was calculated, which provides an indication of the quality of the classifier's confidence judgments. Average precision is calculated the same as the confidence weighted score (see section 10.4) except that the precision is only averaged over the points in the confidence-ranked list where the actual class value is true (entailed). For comparison, Table 13 breaks the results into the subsets

corresponding to the two classifiers. It shows test set results, results from cross-validation on the training sets, the best accuracy for a full submission by anyone at the RTE challenges (Dagan, Glickman and Magnini 2005; Glickman, Dagan and Koppel 2005; Bar-Haim et al. 2006), and the median accuracy of all full submissions.

	SUM/CD	IE	IR	QA	MT	PP	RC	All
RTE1 Accuracy	83.3	49.2	60.0	62.3	63.3	48.0	53.6	61.8
RTE2 Accuracy	70.0	55.5	64.0	55.0	n/a	n/a	n/a	61.1
RTE2 Ave Prec.	80.7	49.4	73.0	57.3	n/a	n/a	n/a	65.2

Table 12. System Accuracy and Average Precision by Task

	SUM / CD	Non-SUM/CD	Overall
RTE1 Training Set CV	83.7	56.9	61.6
RTE1 Test Set	83.3	56.8	61.8
Best RTE1 Submission	83.3	52.8	58.6
Median RTE1 Submission	77.7	49.5	55.2
RTE2 Training Set CV	84.5	62.7	68.1
RTE2 Test Set	70.0	58.2	61.1
Best RTE2 Submission	84.5	72.3	75.4
Median RTE2 Submission	n/a	n/a	58.3

Table 13. System Accuracy by Dataset or Submission

11.3 Discussion

To get a fair comparison with the RTE1 results I trained strictly on the RTE1 training set and tested on the RTE1 test set. As can be seen in Table 13, this system outperformed the submission with the best accuracy at the RTE1 challenge by 3.2%. One RTE1 task, CD, is relatively easy; this system did as well as all but one submission on that task. The non-CD portion of the dataset is very challenging to classify. The best non-CD accuracy submitted to RTE1 was 52.8%, where the accuracy for this system was 4.0% higher at 56.8%. Of the 23 teams in RTE2, this system was among the top ranked systems in both average precision (5^{th}) and accuracy (6^{th}). Only the LCC systems, which trained on three orders of magnitude more data, showed a statistically significant improvement.

12 Experiment One

12.1 Experimental Design

In the first experiment, the intent was simply to provide evidence that a machine learning algorithm could perform reasonably above chance on the task of determining the extent to which elementary school kids understood the low-level facets of the reference answers. The features used were the term co-occurrence maximum likelihood estimates and related features that provided information regarding the statistical validity of the former. Specifically, the MLE was calculated for the facet's governor and modifier, based separately on the surface form of the words and their stems; the product, average and geometric mean of the MLEs over all of the reference answer content words were also included to provide an indication of the learner's overall understanding.

Following the RTE challenge, in this initial investigation, I considered only examples that were given moderately consistent labels by all three annotators (Dagan, Glickman and Magnini 2005, only retained entailment pairs where annotators agreed unanimously and another judge considered the pairs to be reasonable). Specifically, I used the adjudicated labels of facets where all annotators believed the student understood the facet (i.e., labeled the facet Expressed or Inferred), all annotators felt the student contradicted the facet (i.e., labeled the facet Contra-Expr, Contra-Infr or Self-Contra), the adjudicator and at least one annotator chose Diff-Arg, and facets that all annotators labeled Unaddressed. Given the fraction of the dataset that was annotated at the time of

conducting these experiments, this selection resulted in a total of 7273 training examples, 7719 examples in the Unseen Modules test set (the domain-independent test set collected from very different science modules than were included in the training data), 1239 examples in the Unseen Questions test set (answers to questions not used in the training set, but that were from the same science modules) and 424 examples in the Unseen Answers test set (different answers to the same questions that generated the training set answers). Otherwise, I roughly followed the procedure described in chapter 10 to train a classifier on the training examples. In this experiment a Random Forest classifier was utilized, which out of several machine learning algorithms evaluated at that time, achieved the best accuracy in cross validation on the training data.

12.2 Results

Table 14 shows the classifier's accuracy in cross validation on the training set as well as on each of the test sets, Unseen Answers, Unseen Questions, and Unseen Modules. Following the column headings from the corpus annotation, Yes-No presents the accuracy of a two-way classifier that outputs Yes for all facets that the classifier judged as being understood (Expressed or Inferred) and No for all other facets. This effectively is the accuracy of predicting that the tutor should provide some sort of remediation. The column labeled Tutor-Labels provides the accuracy when considering the five classes that will drive the type of dialogue provided by the tutor, Understood (Expressed and Inferred), Contradicted (Contra-Expr and Contra-Infr), Self-Contra, Diff-Arg, and

Unaddressed. The sub-columns first show two simpler baselines, the accuracy of a classifier that always chooses the most frequent class in the training set – Unaddressed – and the accuracy based on a lexical decision that chooses Understood if the stems of both the governing term and the modifier are present in the learner's answer and outputs Unaddressed otherwise. In addition to accuracy, I also calculate the confidence weighted score (CWS), assessing the quality of the classifier's confidence judgments.

	Yes-No Accuracy (%)			Tutor-Labels Acc. (%)			CWS
	Majority Class	Lexical Baseline	Exp. 1 System	Majority Class	Lexical Baseline	Exp. 1 System	
Training Set CV	56.5	65.2	81.9	54.0	62.9	80.3	87.9
Unseen Answers	53.8	64.4	80.9	50.9	61.6	78.8	85.6
Unseen Questions	64.8	73.3	69.7	62.5	71.3	68.4	78.4
Unseen Modules	49.4	72.0	76.6	45.7	68.6	74.6	83.5

Table 14. Exp. 1 Classifier Accuracy and Confidence Weighted Score

12.3 Discussion

This is a new task and new dataset. These early results, based on a very simple lexical similarity approach, are very promising. The results on the Tutor-Labels are 27.9%, 5.9%, and 28.9% over the most frequent label baseline for Unseen Answers, Questions, and Modules respectively. Accuracy on Unseen Answers is 17.2% better than predicting Expressed when both of the facet's word stems are present and Unaddressed in all other cases, and 6% better on Unseen Modules. However, this simpler lexical baseline outperforms a classifier trained

on the more robust lexical features by 2.9% on the Unseen Questions test set. Results on the Yes-No Labels, predicting that the tutor should provide some form of remediation, follow a similar trend. The CWS is much higher than chance, indicating that the confidence (class probability estimates) output by the Random Forest will be useful to the dialog manager in deciding how strongly to believe in the classifier's output. For example, if the classification suggests the learner understood a concept, but the confidence is low, the dialog manager could decide to 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. These confidence estimates could be improved further by following the techniques in (Nielsen 2004). These results demonstrate that the task is feasible and with more rigorous feature engineering, the accuracy should be in a range that allows effective tutoring. Even when the prediction is not correct as long as the tutor acts according to the confidence, no harm and little frustration should result.

13 Experiment Two

13.1 Experimental Design

In this experiment, I added the features described in Table 11 and followed the procedures described in chapter 10. In short, this included the governor and modifier features from experiment one, the subordinate MLEs for inter-propositional facets, a negation consistency check, and the syntactic features: POS, dependency and facet relation types, and the dependency path edit distance. This experiment included all of the data. The entire Unseen Modules dataset was adjudicated, but only about 50% of the training data and other test sets was adjudicated. Again, the facets assumed to be understood a priori were withheld from all of the datasets, resulting in 54967 training examples, 30514 examples in the Unseen Modules test set, 6699 in the Unseen Questions test set and 3159 examples in the Unseen Answers test set. A C4.5 decision tree was trained to classify examples in this experiment, since it performed best in training set cross validation.

13.2 Results

Table 15 shows the classifier's accuracy in cross validation on the training set as well as on each of the test sets. In this experiment, accuracy was only calculated for the labels that will drive the ITS dialogue, *Tutor Labels* – Understood, Contradicted, Self-Contra, Diff-Arg, and Unaddressed. Again, the columns first show two simpler baselines, the accuracy of a classifier that always

chooses the most frequent class in the training set – Unaddressed – and the accuracy based on a lexical decision that chooses Understood if both the governing term and the modifier are present in the learner's answer and outputs Unaddressed otherwise. (I also evaluated placing a threshold on the product of their lexical entailment probabilities similar to Glickman, Dagan and Koppel (2005), who achieved the best results in the first RTE challenge, but this gave virtually the same results as the word matching baseline). The column labeled *Table 11 Features* presents the results of the classifier using all of the additional features. (*Reduced Training* is described in the Discussion section, which follows.)

	Majority Label	Lexical Baseline	Table 11 Features	Reduced Training
Training Set 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	**65.9**

Table 15. Exp. 2 Classifier Accuracy on the Tutor Labels

13.3 Discussion

Including the simple syntactic features and utilizing the full dataset, the results improve over the most frequent class baseline by 24.4%, 3.3%, and 8.0% for Unseen Answers, Questions, and Modules respectively. Accuracy on Unseen Answers is also 19.4% better than the lexical baseline. However, this simple

baseline outperformed the classifier on the Unseen Questions and Unseen Modules test sets.

In the first experiment, the classifier beat the lexical baseline by 6% on the Unseen Modules test set. It seemed probable that the decision tree had over fit the data due to bias in the data itself; specifically, since many of the students' answers are very similar, there are likely to be large clusters of 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 the Unseen Answers test set, but would negatively affect performance on less similar data, such as the Unseen Questions and Modules test sets. To test this hypothesis, I reduced the number of examples in the training set to about 8,000, roughly the number of training examples in the first experiment, and retrained the classifier. This would result in fewer of these dense clusters. The results, shown in the *Reduced Training* column of Table 15, were improvements of 4.8% and 4.5% on the Unseen Questions and Modules test sets, respectively. In the future, I will find a principled means of deciding how much training data and, more specifically, what the make up of the training data should be to optimize generalization to other domains.

13.4 Feature Analysis

Table 4 shows the impact, based on training data cross-validation, of each feature relative to the 54.63% baseline accuracy of always predicting the facet is Unaddressed – the most frequent class in the training set (for this column, the

feature indicated was the only one used by the classifier) and relative to the 77.05% accuracy of a classifier built using all of the features in Table 11 (for this column, the feature indicated was the only one withheld from the learning algorithm's feature set). Positive values in the latter column indicate that the feature hurt the classifier's ability to generalize to held-out cross-validation data.

Feature Added or Removed	Change from Majority Class Baseline (54.63%)	Change from All Features (77.05%)
Gov_MLE	12.97	-0.71
Mod_MLE	12.06	-0.32
Gov_Match	8.10	-0.04
Mod_Match	10.14	+0.03
Gov_Facet's_Gov_MLE	0.67	+0.08
Gov_Facet's_Mod_MLE	0.50	+0.03
Mod_Facet's_Gov_MLE	1.12	+0.12
Mod_Facet's_Mod_MLE	0.91	+0.01
Gov_POS	1.01	-0.55
Mod_POS	1.35	-0.72
Facet_Reltn	1.23	-0.16
AlignedDep_Reltn	-	+0.78
Dep_Path_Edit_Dist	2.97	-0.47
Consistent_Negation	-	-
RA_CW_cnt	3.87	-1.28

Table 16. Feature Impact relative to 1) Baseline and 2) All Features

The most informative features in classifying low-level facet understanding are the lexical similarity features, Gov_MLE and Mod_MLE, derived from our domain-independent co-occurrence statistics. This is consistent with the ability of Latent Semantic Analysis (LSA) to predict understanding in the tutoring environment (Graesser et al. 2001); the difference is that LSA is unable to perform well on the sentence-length answers common in our dataset and, based on earlier investigations, LSA does not process young children's utterances reliably. The exact lexical match features are the second most informative features, but they are redundant with the preceding similarity features (they are *true* when the MLE is precisely 1.0 and *false* otherwise) and, therefore, seem not to add value to the final classifier.

The next four features, the subordinate similarity features, were intended to facilitate the classification of facets that represent relations between other facets or higher-level propositions. They show some marginal value relative to the baseline, but in combination with other features fail to improve the final system's performance. This implies that a more thorough error analysis of this subset of facets must be performed and an alternative approach must be designed for their classification. These same features might still be appropriate if they are used strictly for this relational type of facet.

The facet's dependency relation and the POS tags of the governor and modifier influence how similar aligned dependencies must be to consider them a match; whereas, the label of the aligned dependency appears not to have any predictive value. The benefit from the dependency path edit distance feature

suggests that deeper syntax does help assess understanding at this fine-grained level. Finally, including the number of content words in the reference answer was motivated by the belief that longer answers were more likely to result in spurious alignments; further analysis shows that it is also the case that extended expectations are less likely to be addressed by students.

Table 16 shows that the most salient features are simple lexical features (e.g., lexical co-occurrence statistics). The simple lexical baseline in Table 15 shows an average improvement of about 6.5% on the test sets relative to classifying according to the most frequent class in the training set. Still, the feature analysis shows that syntactic features such as the dependency path edit distance, the facet type, and the POS tags are boosting performance over the purely lexical features.

13.5 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 cross-validation on the training data. Rather than using typical random selection of training examples into *K* data folds, I essentially performed leave-one-out analysis at the module level. In other words, 13 classifiers were built, one for each science module in the training set; each classifier was tested on all of the data in a single science module and trained on the data from the remaining 12 modules. This effectively simulates the Unseen Modules test condition. The feature set shown in Table 11 was utilized to construct each C4.5 decision tree classifier.

13.5.1 Characterizing System Errors

The first consideration was to get a broad characterization of errors, specifically, to determine to what extent system errors are affected by the way reference answers are written and by the type of reference answer facet. The extent to which system accuracy is correlated with human performance is simultaneously checked. The style of questions and reference answers was very different between a number of the science modules, so they represent a logical way to categorize the data for the purpose of assessing whether the way reference answers are written affects system accuracy. In this analysis, only the seven science modules whose answers had been adjudicated are considered (Earth Materials, Ideas and Inventions, Landforms, Levers and Pulleys, Models and Designs, Physics of Sound, and Variables). A chart of the accuracy and ITA for these modules is shown in Fig. 15. I compared the accuracy of each module to the accuracy of all others combined performing a statistical test for the difference of two proportions. With the exception of the two modules nearest the average accuracy, the difference between each module and the data from the remaining six appeared to be statistically significant ($11.1 \leq \chi^2 \leq 45.9$, $p<0.001$, $N=23244$). Next, I compared accuracy across facet relation types in a similar manner. A chart of the accuracy and ITA by relation type is shown in Fig. 16. Again, the test suggested that the accuracy of over two thirds of the relation types was significantly different than the combination of all others ($4.1 \leq \chi^2 \leq 89.9$, $p<0.05$, $N=58126$). As can be seen in the graphs, system accuracy is not, in general, correlated with inter-annotator agreement.

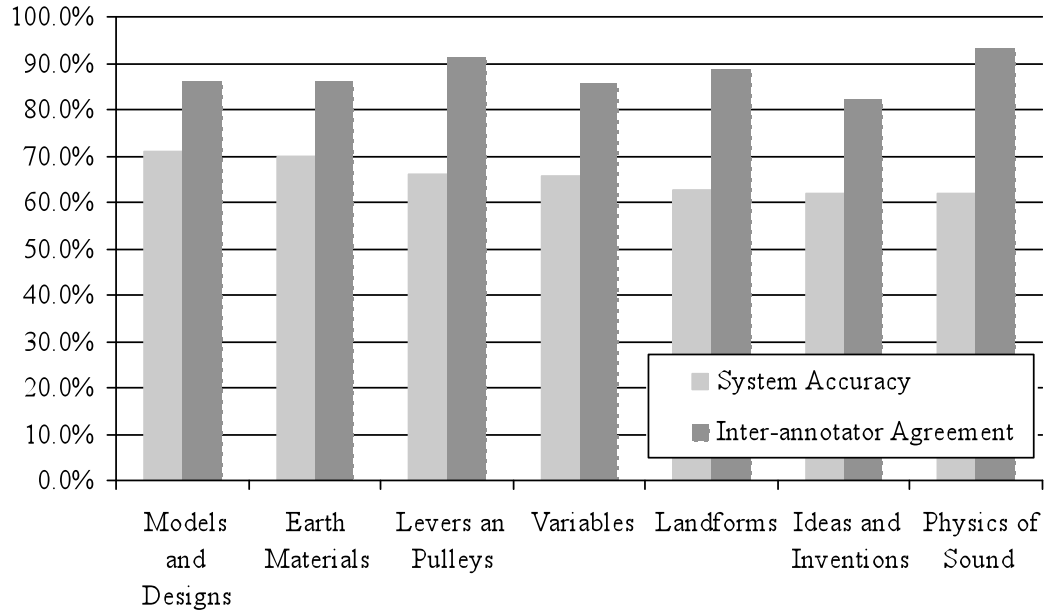

Fig. 15. Classifier Accuracy vs. ITA by Science Module in Decreasing Accuracy

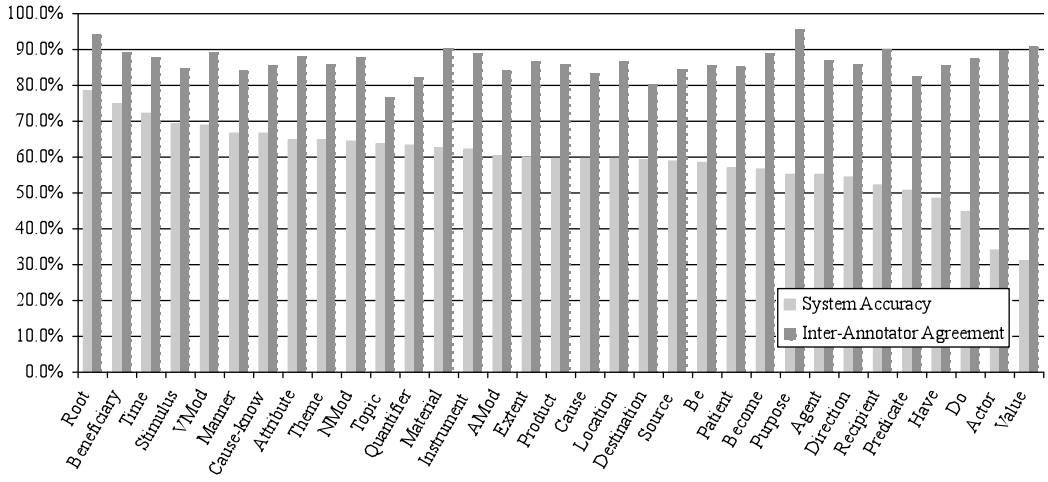

Fig. 16. Classifier Accuracy vs. ITA by Facet Relation Type

While several of the science modules and facet relation types appear to have significantly different characteristics, this should be tempered by the fact that each instance of a reference answer facet has approximately 40 student answers that were annotated relative to it and this dependence between data points

could affect the statistical validity. This is exacerbated by the fact that many of the science modules have questions that are extremely similar leading to identical reference answer facets in extremely similar question contexts. Hence, roughly identical facets could have 80 or perhaps over 160 annotations. Examining just those facets whose type is Agent, we see that the characteristics of individual instances of reference answer facets themselves vary significantly from 7.5% to 87.5% accuracy – the 20 facets with the best accuracy and the 20 with the worst accuracy are each statistical different from the collection of all other facets ($5.0 \leq \chi^2 \leq 37.9$, $p<0.05$, $N=2850$). A chart of the average accuracy associated with each of the 72 Agent type reference answer facets is shown in Fig. 17.

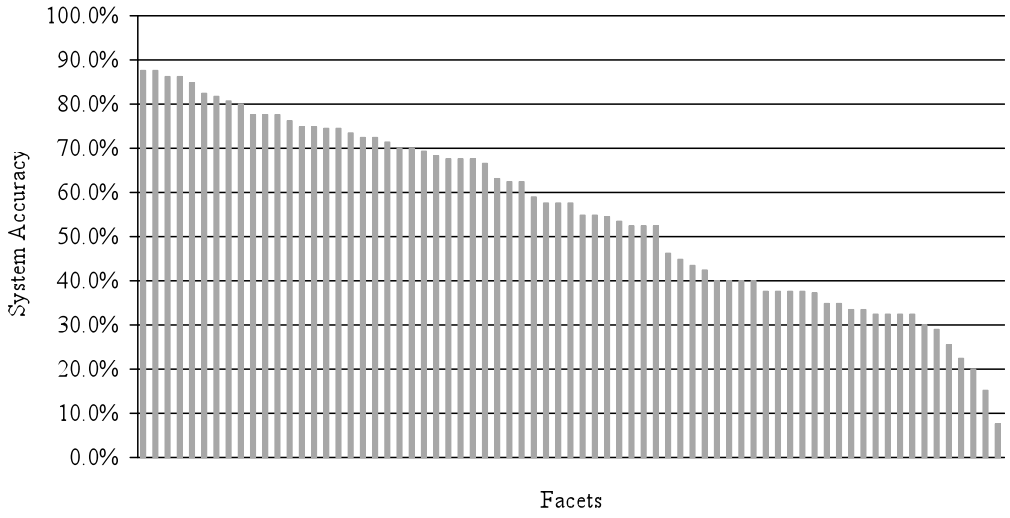

Fig. 17. System Accuracy for Agent Reference Facets, ~40 Annotations Each

13.5.2 Errors in Expressed Facets

Next, several randomly selected examples were analyzed to look for patterns in the types of errors the system makes. However, only specific categories of data were considered. Specifically, only the subsets of errors that

were most likely to lead to short-term system improvements were considered. This included only examples where all of the annotators agreed on the annotation, since if the annotation was difficult for humans, it would probably be harder to construct features that would allow the machine learning algorithm to correct its error (although the lack of correlation between the system accuracy and ITA as seen in Fig. 15 and Fig. 16 suggests this might not be true). Second, only Expressed and Unaddressed facets were considered, since Inferred facets represent the more challenging judgments, typically based on pragmatic inferences. Contradictions were excluded since there was no attempt to handle these in the present system. Third, only facets that were not inter-propositional were considered, since the inter-propositional facets are more complicated to process and only represent 12% of the non-Assumed data. Expressed facets are discussed in this section of the thesis and Unaddressed in the next section.

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 facet. I analyzed 100 random examples of errors where annotators considered the facet Expressed, but the analysis only considered one example for any given reference answer facet. Out of these 100 errors, only one looked as if it was probably incorrectly annotated. The potential error factors seen in the data included synonymy, hypernymy, hyponymy, meronymy, derivational changes, other lexical paraphrases, concept definitions, pronoun

resolution, noun phrase term substitution, other coreference resolution, negation, syntactic transformations, dependency parser problems, semantic role parsing, paraphrasing, logical or deep reasoning, pragmatics, and corpus-related issues. In over half of the examples, there were two or more factors involved (this is probably an underestimate, due to focusing on the most significant or obvious factors involved in a given example).

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 35 occurrences. Roughly half of these relationships should be detectable using a broad coverage lexical resource. For example, substituting *tiny* for *small*, *CO2* for *gas*, *put* for *place*, *pen* for *ink* and *push* for *carry* (WordNet entailment). However, many of these lexical paraphrases are not necessarily associated in lexical resources such as WordNet. For example, in the substitution of *put the pennies* for *distribute the pennies*, these terms are only connected at the top of the WordNet hierarchy at the Synset (*move, displace*). Similarly, WordNet appears not to have any connection at all between *have* and *contain*. Concept definitions account for an additional 14 issues that could potentially be addressed by a lexical resource such as WordNet.

The three coreference resolution factors combined are involved in nearly 30% of the errors. Students use on average 1.1 pronouns per answer and, more importantly, the pronouns tend to refer to key entities or concepts in the question and reference answer. A pronoun was used in 15 of the errors (3 personal

pronouns – *she*, 11 uses of *it*, and 1 use of *one*). It might be possible to correct many of these errors by simply aligning the pronouns to essentially all possible nouns in the reference answer and then choosing the alignment that gives the learner the most credit. However, confidence in the benefit of pronoun resolution has to be tempered by the fact that pronouns occurred just as frequently in a random sample of 300 examples that the system correctly classified as they occurred in 300 incorrectly classified examples. In 6 errors, the student referred to a concept by another term (e.g., substituting *stuff* for *pieces*). In 6 errors, the student used one of the terms in a noun phrase to refer to a concept where the reference answer facet included the other term as its modifier or vice versa. For example, one reference answer was looking for NMod(*particles, clay*) and Be(*particles, light*) and the student said *Because clay is the lightest*, which should have resulted in an Understood classification for the second facet. This type of error should be easily overcome by adding features that generally allow the substitution of any term in the noun phrase (e.g., features that output the similarity for the best matching word in the noun phrase).

The three most common issues (both in isolation and in combination with other factors) were deep or logical reasoning, pragmatics and phrase-based paraphrasing. At least one of these factors is involved in almost two thirds of the errors. Examples of the first issue include recognizing that both cups have the same amount of water given the following student response, *no, cup 1 would be a plastic cup 25 ml water and cup 2 paper cup 25 ml and 10 g sugar*, and that two sounds must be *very different* in the case that *... it is easy to discriminate...*

Examples of pragmatic issues include recognizing that saying *Because the vibrations* implies that a rubber band is vibrating given the question context, and that *the earth* in the response ... *the fulcrum is too close to the earth* should be considered to be *the load* referred to in its reference answer. These two factors, pragmatics and logical reasoning, were involved in nearly 40% of the errors. It is interesting that these are all examples that three annotators unanimously considered to be Expressed versus Inferred facets. Many of the remaining errors were classified as involving phrase-based paraphrases. Examples here include ... *it will heat up faster* versus *it got hotter faster* and *in the middle* versus *halfway between*. This is an area that I intend to invest significant time in future research. This research should also reduce the error rate on lexical paraphrases.

Of the remaining errors, corpus issues and negation are the only categories that occurred in isolation, all others were seen in combination with other factors. At least two errors appeared to result from corpus-related issues, the system being unable to resolve *3* to *three* and *g* to *grams*. Many of these types of terms should be normalized in the corpus since the automatic speech recognition will output a consistent form, but it is debatable whether everything should be normalized. Students are unlikely to say *5 g* so converting this to *5 grams* is appropriate. However, some students are likely to say *CO2*, so converting this to *carbon dioxide* is more questionable. For domain specific tutoring, this can be handled as a preprocessing step, but to handle new questions or topics, the system should be able to resolve these terms. Six errors essentially involved negation of an antonym, (e.g., substituting *not a lot* for *little* and *no one has the same fingerprint*

for *everyone has a different print*). While syntactic variation is certainly common in the data, it did not appear to be the primary factor in any of the errors. Semantic role labeling has the potential to provide the classifier with information that would clearly indicate the relationships between the student and the reference answer, but there was only one error in which this came to mind as an important factor and it was not due to the role labels themselves, but because MaltParser identifies only a single head. Specifically, in the sentence *She could sit by the clothes and check every hour if one is dry or not*, the pronoun *She* is attached as the subject of *could* (*sit*), but *check* is left without a subject.

Errors in the dependency parses seemed likely to be contributing to the system error rate. In previous work, analyzing the dependency parses of fifty one of the student answers, many had what were believed to be minor errors, 31% had significant errors, and 24% had errors that looked like they could easily lead to problems for the answer assessment classifier. Over half of the more serious dependency parse errors resulted from inopportune sentence segmentation due to run on student sentences conjoined by *and* (e.g., the parse of *a shorter string makes a higher pitch and a longer string makes a lower pitch*, errantly conjoined *a higher pitch* and *a longer string* as the subject of *makes a lower pitch*, leaving *a shorter string makes* without an object). To overcome these issues, the text could be parsed once using the original sentence segmentation and then again with alternative segmentations under conditions to be determined by further dependency parser error analysis. One partial approach could be to split the sentence when two noun phrases are conjoined and they occur between two verbs,

as is the case in the preceding example, where the alternative segmentation resulted in correct parses. Then the system could choose the parse that is most consistent with the reference answer. While I believe improving the parser output will result in higher accuracy by the classifier, there was little evidence to support this in the answer assessment system error analysis. I only checked the parses when it was somewhat surprising that the classifier made an error (for example, when there were simple lexical substitutions involving very similar words) and the dependency path features looked wrong. Only two of these classification errors were associated with parser errors. Still, I believe that better parses should lead to more reliable (less noisy) features, which in turn will allow the machine learning algorithm to more easily recognize what is important.

Finally, many of the dependency path features were completely different than what they should have been if the lexical alignment was right. Given that the majority of parses were correct in the areas relevant to the analyzed reference answer facets, it is likely that the alignment is wrong. Future plans include training an alignment classifier separate from the assessment classifier. This will at minimum facilitate the analysis of alignments, which are key to understanding the learner answers.

In closing, it should be emphasized that over half of the errors in Expressed facets involved more than one of the 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. Note, I think that question was kind of dumb*,

requires resolving *it* to the tree and the *solid object* mentioned to the tree, and then making the inference that *light cannot go through [the tree]* entails the tree blocks the light.

13.5.3 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 fragment *You could take a couple of cardboard houses and ... 1 with thick glazed insulation. ...*, all three annotators suggested they could not infer the student meant the insulation should be installed in one of the houses. 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*. Stating a function of the elytra was *horn* was insufficient for the annotators to credit the student with understanding that it was used to *make sounds*.

One of the biggest sources of errors in Unaddressed facets is the result of ignoring the context of words. For example, consider the question *When you make an electromagnet, why does the core have to be iron or steel?* and its reference answer *Iron is the only common metal that can become a temporary magnet. Steel is made from iron.* Then, given the student answer *It has to be iron or steel because it has to pick up the washers*, the system classified the facet Material_from(*made, iron*) as Understood based on the text *has to be iron*, but ignores the context, specifically, that this should be associated with a production, Product(*made, steel*). Similarly, 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.

A fair number of the errors in Unaddressed facets appear to be the result of antonyms having very similar statistical co-occurrence patterns. Examples of errors here include confusing *closer* with *greater distance* and *absorbs energy* with *reflects energy*. However, both of these also may be annotation errors that should have been labeled Contra-Expr.

The biggest source of error is simply classifying a number of facets as Understood if there is some lexical similarity and at times some syntactic similarity as in the case of accepting *the balls are different* in place of *different girls*. However, there are also a fair number of cases where it is unclear why the decision was made, as in the following case, where the system apparently trusts that the student understands a complicated electrical circuit based on the answer *I learned it in class*.

My belief is that with the processes and the more informative features described in the previous subsection, the learning algorithm will 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.

14 Discussion and Future Work

Table 16 shows that the most salient machine learning features are simple lexical features (e.g., co-occurrence statistics). The simple lexical baseline shows an average improvement of about 6.5% relative to classifying according to the most frequent class in the training set (see Table 15). Still, error analysis suggests that additional features related to lexical similarity could boost performance substantially. Many of these features will be extracted from lexical resources such as WordNet.

Additional lexical relatedness features that will be considered include the Jiang-Conrath distance, the Extended Lesk measure, and Latent Semantic Analysis. The Jiang-Conrath distance measures the distance between words using Information Content; the distance between two words is the amount of information required to represent their commonality minus the information needed to describe both words (Jiang and Conrath 1997). Budanitsky and Hirst (2006), as well as many other researchers, have found the Jiang-Conrath distance to be the best measure of semantic relatedness they tested in their evaluation framework (they did not test the Extended Lesk measure). The Extended Lesk measure provides a measure of the overlap between the glosses of each word and between the glosses of their various relations such as hyponyms and hypernyms (Banerjee and Pedersen 2003). Banerjee and Pedersen show a slight advantage to this metric in their evaluation framework. These metrics are expected to be good measures of semantic relatedness and, therefore, good alignment metrics, but they are not expected to be useful in distinguishing entailing relations from

contradictory relations. That is a desirable feature in this case, since we not only want to recognize entailments, but also contradictions. For example, given the reference answer *The pitch rises*, if the learner says *The pitch falls*, we would like to consider *falls* to be a good alignment with *rises*. At this stage in the processing we want to detect all potentially related terms and then, in a later stage, determine whether the dependency and propositional relationship is one of entailment, contradiction, or neither.

In addition to the co-occurrence and lexical resource features, I will be generating collocation features to provide evidence for when two words can be used in similar contexts. These features indicate the extent to which two words tend to co-occur with shared governors, modifiers, and other context words.

The lexical metrics are utilized to sort the list of learner terms by relatedness to a given reference answer term and to filter out terms assumed to be unrelated due to a poor matching score. Error analysis suggests that additional information be extracted from the system in order to analyze this alignment in more detail and, most likely, I should follow through on plans to split the task into two classification steps, first performing an alignment and then assessing the learner answer.

The current feature set is really intended strictly to differentiate between those facets of the reference answer that the student most likely understands from those that were not addressed. This leaves the approach prone to some of the same problems discussed earlier for LSA, (e.g., it does not explicitly distinguish antonyms from synonyms). This must be addressed in future work, since despite

their somewhat infrequent occurrence, it is critical that the automated tutor recognize and address these contradictory beliefs as early as possible. The most likely technique for recognizing antonyms is to use a broad coverage thesaurus. The system might also benefit from including versions of the PMI-IR metric that consider the context of the co-occurring terms and that are intended to differentiate contradictions (e.g., antonyms) from paraphrases and synonyms.

One of my highest priority areas of future research is implementing a paraphrase detection module based on the work described in section 4.2. This module will check whether there is evidence in a large corpus to suggest that based on lexical and dependency relations a phrase in the learner answer is a paraphrase of part of the reference answer. However, the error analysis implies that existing paraphrase recognition techniques will at minimum require significant modification.

Many of the learning algorithm's features, both implemented and planned, rely on extracting statistics from large corpora. Currently, these corpora are virtually all drawn from the news domain. Since the vocabulary is quite different than that used in elementary school science, this undoubtedly has a significant negative effect on system performance. Several additional, more relevant resources must be collected and indexed.

The current feature set was largely constructed with domain-independent assessment in mind. Several additional features could easily improve the accuracy of question-dependent assessment. For example, simply adding unique question and facet identifiers would allow the learning algorithm to associate the

relevance of specific feature values with individual contexts. A related immediate area of research to improve results on the domain-independent Unseen Modules test set includes investigating techniques to avoid over-fitting the classifier to the same facet-specific characteristics we want to capitalize on above.

Other short-term areas of research include coreference resolution and improving the dependency parser performance. Coreference was one of the most frequent issues seen in the error analysis and has a potentially easy fix in selecting the alignment most consistent with the reference answer semantics. While it was not as clear that dependency parser errors were directly causing assessment errors, several dependency paths looked suspect and I believe that less noisy features might allow the machine learning algorithm to detect patterns it is currently missing. The primary initial effort here is simply in revising the current sentence segmentation.

The current system relies on hand-generated reference answer facets. To consider this a truly domain-independent approach, these facets must be extracted by an automated parser. Further research is also required to determine what factors are most important in constructing reference answer facets and exactly how text should be restructured in order to facilitate this high-value facet extraction.

When integrating this semantic assessment module in the eventual tutoring system, probabilistic reasoning should be used to decide whether and how to address apparent contradictions, misconceptions and unaddressed issues on the part of the student. In the first experiment here, confidence weighted scores

approximately 10% (absolute) over the classification accuracy were achieved, indicating that the class probability estimates will be useful to the dialog manager in deciding how strongly to believe in the classifier's output. For example, if the classification suggests the learner understood a concept, but the confidence is low, the dialog manager could decide to paraphrase the answer as a transition to the next question, rather than assuming the learner definitely understands and simply moving on or rather than asking a confirming question about something the learner probably already expressed. Additional research is necessary to achieve better probability estimates, but more importantly to decide exactly how to use these confidence measures within the dialog management.

Of course the most important, and perhaps biggest, area of future research involves the integration of this assessment technology into the ITS. There are both short-term research issues involving how to combine this system's output with other assessment techniques and long term issues associated with utilizing the output to the best possible advantage in driving the tutoring dialog to optimize student learning gains.

15 Conclusions and Broader Impact

The three most significant contributions of this work are 1) formally defining and evaluating a representation and learner answer classification scheme that involves the annotation of detailed answer facets with the fine-grained classifications necessary to enable more intelligent out-of-domain dialog control, 2) laying the framework for a domain-independent answer assessment system that can classify learner responses to previously unseen questions according to this scheme, and 3) the creation of a public corpus of student answers annotated according to this method. This work will facilitate the creation of an effective and scalable tutoring system that represents a significant advance over the state of the art.

In chapters 6 through 9, I presented a case for the benefit of more detailed learner answer assessment than has been attempted in prior work. I provided evidence for this benefit in the analysis of a number of specific learner answers and described a knowledge representation and annotation scheme that would support such an assessment. The corpus of learner answers described here was annotated with substantial agreement (86.1%, Kappa = 0.728) and will be made publicly available for other researchers to utilize in improving their tutoring and educational assessment technologies. There are currently no publicly available corpora of learner answers that researchers can utilize for these purposes. This database of annotated answers provides a shared resource and a standardized annotation scheme allowing researchers to compare work and should stimulate further research in these areas.

This labeled corpus and the associated representation and annotation scheme is also expected to result in important advances in the state-of-the-art in textual semantic entailment. This is essential in a wide variety of applications outside of intelligent tutoring systems, such as question answering (Harabagiu and Hickl 2006), information extraction, machine translation, machine reading, and many others. Most current techniques as demonstrated by the Pascal Recognizing Textual Entailment challenges do not perform much above chance.

In chapters 10 through 14, I presented an approach to automatically assess learner answers utilizing the novel representation and assessment scheme described in the first half of this thesis. The results presented here for the Unseen Answers test set are 24.4% better than the majority class baseline and 19.4% better than a baseline derived from the best performing system at the first RTE challenge. This demonstrates that the basic within-domain classification task is feasible and with more rigorous feature engineering, accuracy will easily be in a range that allows effective tutoring. The out-of-domain results are 12.5% and 3% better than the most frequent class and lexical baselines respectively. This represents reasonable performance for an initial domain-independent system – the first RTE systems all failed to outperform this same lexical baseline.

Even when the prediction is not correct, as long as the tutor acts according to the confidence, the dialog can be effective. Accurate probabilities will allow the dialog manager to decide whether to, for example, assume a misconception with high confidence and take appropriate corrective action or decide, due to low confidence, that it should clarify the learner's understanding on a particular facet

of the question. Prior work in the area of tutoring and answer verification has not explicitly addressed the benefits of probabilistic outputs.

To my knowledge, this is the first work to demonstrate success in assessing roughly sentence-length constructed answers from elementary school children. These are the kind of responses that might be expected in an inquiry-based tutoring environment. Improving reading and science comprehension in these formative years is critical in order to establish a foundation for later learning.

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.

This work represents an important prerequisite to achieving the goal of significantly increasing the learning gains effected by intelligent tutoring systems. I hope that the end result will stimulate additional research into the kind of individualized tutoring envisaged below by Bennet:

With intelligent tutors particularly, student knowledge will be dynamically modeled using cognitive and statistical approaches capable both of guiding instruction on a highly detailed level and of providing a general summary of overall standing. Instruction will be adapted not only to the multiple dimensions that characterize standing in a broad skill area, but to personal interests and background, allowing more meaningful accommodation to diversity than was possible with earlier approaches.

Bennet, R. 1998, Reinventing Assessment: Speculations on the Future of Large-Scale Educational Testing

References

Agichtein, E and Gravano, L. (2000). Snowball: Extracting relations from large plain-text collections. In *Proc. of the 5th ACM International Conference on Digital Libraries*.

Akhmatova, E. (2005). Textual Entailment Resolution via Atomic Propositions. In *Proc. of the PASCAL Recognizing Textual Entailment Challenge Workshop*.

Aleven, V., Popescu, O., and Koedinger, K.R. (2001). A tutorial dialogue system with knowledge-based understanding and classification of student explanations. In *Working Notes of 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

Anderson, J.R., Corbett, A.T., Koedinger, K., and Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of Learning Sciences*, 4, 167-207.

Andreevskaia, A., Li, Z., and Bergler, S. (2005). Can Shallow Predicate Argument Structures Determine Entailment? In *Proceedings of the PASCAL Recognizing Textual Entailment Challenge Workshop*.

Bach, E. (1986). The algebra of events. *Linguistics and Philosophy* 9: 5–16.

Banerjee, S. and Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI 2003*, 805–810.

Bar-Haim, R, Szpektor, I, and Glickman, O. (2005). Definition and Analysis of Intermediate Entailment Levels. In *Proc. of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, 55-60.

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B. and Szpektor, I. 2006. The Second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

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

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

Bayer, S., Burger, J., Ferro, L., Henderson, J., and Yeh, A. (2005). MITRE's Submissions to the EU Pascal RTE Challenge. In *Proceedings of the PASCAL Recognizing Textual Entailment Challenge Workshop*.

Beck, I.L., McKeown, M.G., Worthy, J., Sandora, C.A., and Kucan, L. (1996) Questioning the author: A year-long classroom implementation to engage students with text. In *The Elementary School Journal, 96*(4), 387-416.

Bejan, C.A., Moschitti, A., Morărescu, P., Nicolae, G., and Harabagiu, S. (2004). Semantic parsing based on FrameNet. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain, July 2004.

Bennett, R. (1998). *Reinventing Assessment: Speculations on the Future of Large-Scale Educational Testing*. Educational Testing Service. Downloaded from http://www.ets.org/research/pic/bennett.html, on Aug. 16, 2005.

Bethard, S., Nielsen, R.D., Martin, J.H., Ward, W., and Palmer, M.. (2007). Semantic Integration in Learning from Text. In *Proc. Machine Reading AAAI Spring Symposium*.

Bloom, B. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive Domain*. New York: McKay.

Bloom, B.S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring. *Educational Researcher 13*, 4-16.

Blum, A. and Mitchell, T. (1998). Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of Computational Learning Theory '98*.

Boonthum, C. (2004). iSTART: Paraphrase Recognition. In *Proc. of the Student Research Workshop, ACL*.

Bos, J. and Markert, K. (2005). Combining Shallow and Deep NLP Methods for Recognizing Textual Entailment. In *Proceedings of the PASCAL Recognizing Textual Entailment Challenge Workshop*.

Brants, T., and Franz, A. (2006). *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia

Braz, R.S., Girju, R., Punyakanok, V., Roth, D., and Sammons, M. (2005). An Inference Model for Semantic Entailment in Natural Language. In *Proc. of the PASCAL Recognizing Textual Entailment Challenge Workshop*.

Breiman, L. (2001). Random Forests. *Journal of Machine Learning*, 45(1):5–32.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. In *Computational Linguistics*, 32(1).

Burger, J. and Ferro, L. (2005). Generating an Entailment Corpus from News Headlines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, 49–54.

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

Carlson, L., Marcu, D., and Okurowski, M.E.. (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech*, Denmark.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of NAACL*, 132–139, Seattle, Washington.

Chi, M.T.H. (1996). Constructing Self-Explanations and Scaffolded Explanations in Tutoring. In *Applied Cognitive Psychology, Vol. 10*, S33–49.

Church, K.W. and Hanks, P. (1989). Word association norms, mutual information, and lexicography. In *Proceedings of the 27th ACL*, Vancouver, B.C., 76–83. ACL.

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

Cohen, P.A., Kulik, J.A., and Kulik, C.L.C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*, 237-248.

Cole, R., Van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W. and Yan, J. (2003). Perceptive Animated Interfaces: First Steps Toward a New Paradigm for Human Computer Interaction.

Cowie, J., Lehnert, W.G. (1996). Information Extraction. In *Communications of the ACM, 39*(1), 80-91.

Dagan, I., Glickman, O., and Magnini, B. (2005). The PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the PASCAL Recognizing Textual Entailment Challenge Workshop*.

Delmonte, R., Tonelli, S., Boniforti, M.A.P., Bristot, A., and Pianta, E. (2005). VENSES – a Linguistically-Based System for Semantic Evaluation. In *Proceedings of the PASCAL Recognizing Textual Entailment Challenge Workshop*.

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

Duclaye, F., Yvon, F. and Collin, O. (2002). Using the Web as a Linguistic Resource for Learning Reformulations Automatically. *LREC*.

Fellbaum, C. (1998). *WordNet An Electronic Lexical Database*. Edited by Christiane Fellbaum. MIT Press.

Fowler, A., Hauser, B., Hodges, D., Niles, I., Novischi, A., and Stephan, J. (2005). Applying COGEX to Recognize Textual Entailment. In *Proceedings of the PASCAL Recognizing Textual Entailment Challenge Workshop*.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., and Dooley, S. (2005) Summary Street: Computer support for comprehension and writing. *Journal of Educational Computing Research*, 33:53-80.

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

Glickman, O. and Dagan, I. (2003). Identifying lexical paraphrases from a single corpus: A case study for verbs. *Recent Advantages in Natural Language Processing (RANLP-03)*.

Glickman, O. and Dagan, I., and Koppel, M. (2005). Web Based Probabilistic Textual Entailment. In *Proceedings of the PASCAL Recognizing Textual Entailment Challenge Workshop*.

Gomez, K., Kwon, S., Gomez, L., and Sherer, J. (In press) Supporting Reading-to-Learn in Science: The Application of Summarization Technology in Multicultural Urban High School Classrooms. In R, Bloymeyer, T. Ganesh, and H. Waxman (Eds.) *Research in Technology Use in Multicultural Settings*. Charlotte, NC: Information Age Publications.

Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., and the Tutoring Research Group. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments, 8*(2):87-109.

Graff, D. (2003). English Gigaword. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05

Graesser, A.C., Hu, X., Susarla, S., Harter, D., Person, N.K., Louwerse, M., Olde, B., and the Tutoring Research Group. (2001). AutoTutor: An Intelligent Tutor and Conversational Tutoring Scaffold. In *Proceedings for the 10th International Conference of Artificial Intelligence in Education* San Antonio, TX, 47-49.

Grice, H.P. (1975). Logic and conversation. In P Cole and J Morgan, editors, *Syntax and Semantics, Vol 3, Speech Acts*, 43–58. New York: Academic Press.

Haghighi, A.D., Ng, A.Y., and Manning, C.D. (2005). Robust Textual Inference via Graph Matching. In *Proc. HLT-EMNLP*.

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

Herrera, J., Peñas, A., and Verdejo, F. (2005). Textual Entailment Recognition Based on Dependency Analysis and WordNet. In *Proc. of the PASCAL Recognizing Textual Entailment Challenge Workshop*.

Hickl, A., Bensley, J., Williams, J., Roberts, K., Rink, B., and Shi, Y. (2006). Recognizing Textual Entailment with LCC's GROUNDHOG System. In *Proc. of the Second PASCAL Recognizing Textual Entailment Challenge Workshop*.

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*

Hyvärinen, A. and Erkki, O. (2000). Independent Component Analysis: Algorithms and Applications. In *Neural Networks*, 13(4-5):411-430.

Jiang, J.J. and Conrath, D.W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics (ROCLING X)*, pages 19–33, Taiwan.

Jordan, P.W., Makatchev, M., and VanLehn, K. (2004). Combining competing language understanding approaches in an intelligent tutoring system. In J. C. Lester, R. M. Vicari, and F. Paraguacu, (Eds.), *7th Conference on Intelligent Tutoring Systems*, 346-357. Springer-Verlag Berlin Heidelberg.

Jurafsky, D. and Martin, J.H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice-Hall. Chapter 14.

Kilgarriff, A. and Rosenzweig, J. (2000). Framework and results for English SENSEVAL. *Computers and the Humanities*, 34(1-2).

Kingsbury, P., Palmer, M., and Marcus, M. (2002). Adding semantic annotation to the Penn Treebank. In *Proceedings of the HLT-02*.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, MA: Cambridge University Press.

Kipper, K., Dang, H.T., and Palmer, M. (2000). Class-Based Construction of a Verb Lexicon. *AAAI Seventeenth National Conference on Artificial Intelligence*, Austin, TX.

Koedinger, K.R., Anderson, J.R., Hadley, W.H., and Mark, M . A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8*, 30-43.

Kohavi, R. (1995) The power of decision tables. In *Proc. of the eighth European Conference on Machine Learning*, pp 174–189. Springer.

Landauer, T.K., and Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*.

Landauer, T.K., Foltz, P.W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

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. (2004). Scoring free-response automatically: A case study of a large-scale Assessment. *Examens, 1*(3).

Leacock, C. and Chodorow, M. (2003). C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities, 37*:4.

Levinson, S.C. (1983). *Pragmatics*. Cambridge University Press, UK.

Lewis, D.D., Yang, Y., Rose, T., and Li, F. (2004). RCV1: A New Benchmark Collection for Text Categorization Research. *JMLR*, 5:361-397.

Lin, D. (1993). Principle-Based Parsing Without OverGeneration. In *Proceedings of ACL-93*, 112-120. Columbus, OH.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 898–904, Montreal.

Lin, D. and Pantel, P. (2001a). DIRT – Discovery of inference rules from text. In *Proc. of KDD*.

Lin, D. and Pantel, P. (2001b). Discovery of inference rules for Question Answering. In *Natural Language Engineering, 7*(4):343-360.

MacCartney, B., Grenager, T., de Marneffe, M., Cer, D., and Manning, C. (2006). Learning to recognize features of valid textual entailments. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006)*.

Madden, N.A., and Slavin, R.E. (1989). Effective pullout programs for students at risk. In *Effective Programs for Students At Risk*, R.E. Slavin, N. L. Karweit, and N.A. Madden, eds. Boston: Allyn and Bacon.

Makatchev, M., Jordan, P., and VanLehn, K. (2004). Abductive Theorem Proving for Analyzing Student Explanations and Guiding Feedback in Intelligent Tutoring Systems. *Journal of Automated Reasoning for Special Issue on Automated Reasoning and Theorem Proving in Education*. 32(3), 187-226.

Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The Penn TreeBank: Annotating predicate argument structure.

Mathews, E.C., Jackson, G.T., Person, N.K., and Graesser, A.C. (2003). Discourse Patterns in Why/AutoTutor. *Proceedings of the 2003 AAAI Spring Symposia on Natural Language Generation*, 45–51. Palo Alto, CA: AAAI Press.

Meichenbaum, D. and Biemiller, A. (1998). *Nurturing independent learners: Helping students take charge of their learning*. Cambridge, MA: Brookline.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B. and Grishman, R. (2004). The NomBank Project: An Interim Report. In *Proc. HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, 24–31.

Miltsakaki, E., Prasad, R., Joshi, A. and Webber, B. (2004). The Penn Discourse TreeBank. In *Proceedings of the Language Resources and Evaluation Conference*, Lisbon, Portugal.

Mitchell, T., Aldridge, N., and Broomhead, P. (2003). Computerized marking of short-answer free-text responses. *Paper presented at the 29th annual conference of the International Association for Educational Assessment (IAEA)*, Manchester, UK.

Mitchell, T., Russell, T., Broomhead, P. and Aldridge, N. (2002). Towards Robust Computerized Marking of Free-Text Responses. In *Proc. of 6th International Computer Aided Assessment Conference*, Loughborough.

Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., Huang, C., Junker, B., Sklar, M.B., and Tobin, B. (2003). Evaluation of an automated Reading Tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research, 29*(1), January, 2003, 61 - 117.

NAEP, (2005, 2007). http://nces.ed.gov/nationsreportcard/

Nielsen, R.D. (2004). MOB-ESP and other Improvements in Probability Estimation. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*.

Nielsen, R.D. and Pradhan, S. (2004). Mixing Weak Learners in Semantic Parsing. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, Barcelona, Spain, July 25-26.

Nielsen, R.D., Ward, W., and Martin, J.H. (2006). Toward Dependency Path based Entailment. In *Proc. of the second PASCAL Recognizing Textual Entailment Challenge Workshop*.

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

Nielsen, R.D., Ward, W., Martin, J.H. (2007) Soft computing in Intelligent Tutoring Systems and Educational Assessment. In *Soft Computing Applications in Business*. Springer.

Nigam, K. and Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*.

Nivre, J. and Scholz, M. (2004). Deterministic Dependency Parsing of English Text. In *Proceedings of COLING*, Geneva, Switzerland, August 23-27.

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

Nivre, J., Hall, J., Nilsson, J., Eryigit, G. and Marinov, S. (2006). Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL)*.

Nivre, J., and Kubler, S. (2006). Dependency Parsing, *Tutorial at COLING-ACL*, Sydney, Australia.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. In *Computational Linguistics*.

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

Pazienza, M.T., Pennacchiotti, M., and Zanzotto, F.M. (2005). Textual Entailment as Syntactic Graph Distance: a rule based and a SVM based approach. In *Proc. of the PASCAL Recognizing Textual Entailment Challenge Workshop*.

Pellom, B. (2001) "SONIC: The University of Colorado Continuous Speech Recognizer", *University of Colorado, tech report #TR-CSLR-2001-01*, Boulder, Colorado, March.

Pellom, B., Hacioglu, K. (2003) "Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task", in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, April.

Peters, S., Bratt, E.O., Clark, B., Pon-Barry, H., and Schultz, K. (2004). Intelligent Systems for Training Damage Control Assistants. In *Proc. of Interservice/Industry Training, Simulation, and Education Conference*.

Platt, J. 2000. Probabilities for Support Vector Machines. In A. Smola, P. Bartlett, B. Scolkopf, and D. Schuurmans (Eds), *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA.

Pon-Barry, H., Clark, B., Schultz, K., Bratt, E.O. and Peters, S. (2004). Contextualizing Learning in a Reflective Conversational Tutor. In *Proceedings of the 4th IEEE International Conference on Advanced Learning Technologies*.

Pradhan, S., Ward, W., Hacioglu, K., Martin, J.H., and Jurafsky, D. (2005). Semantic Role Labeling Using Different Syntactic Views. In *Proceedings of ACL*.

Pulman, S.G. and Sukkarieh, J.Z. (2005). Automatic Short Answer Marking. In *Proc. of the 2nd Workshop on Building Educational Applications Using NLP, ACL*.

Quinlan, J.R. (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, M.C., Manning, C.D., and Ng, A.Y. (2005). Robust Textual Inference using Diverse Knowledge Sources. In *Proc. of the PASCAL Recognizing Textual Entailment Challenge Workshop*.

Ravichandran, D. and Hovy, E. (2002). Learning Surface Text Patterns for a Question Answering system. In *Proc. of the 40th ACL conference*. Philadelphia, PA.

Roll, I., Baker, R.S., Aleven, V., McLaren, B.M., and Koedinger, K.R. (2005). Modeling Students' Metacognitive Errors in Two Intelligent Tutoring Systems. In L. Ardissono, P. Brna, and A. Mitrovic (Eds.), *User Modeling*, 379–388.

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

Rosé, C.P., Gaydos, A., Hall, B.S., Roque, A., and VanLehn, K. (2003b). Overcoming the Knowledge Engineering Bottleneck for Understanding Student Language Input. *Proceedings of AI in Education*. Amsterdam: IOS Press.

Shaw, S. (2004). Automated writing assessment: a review of four conceptual models. In *Research Notes, Cambridge ESOL*. Downloaded from http://www.cambridgeesol.org/rs_notes/rs_nts17.pdf August 10, 2005.

Shinyama, Y. and Sekine, S. (2003). Paraphrase Acquisition for Information Extraction. In *Proc. of The Second International Workshop on Paraphrasing (IWP2003)*, Sapporo, Japan.

Shinyama, Y., Sekine, S., Sudo, K., and Grishman, R. (2002). Automatic paraphrase acquisition from news articles. In *Proc. HLT*. San Diego, California.

Snow, C. (2002). *Reading for Understanding: Toward A R & D Program in Reading Comprehension*, Rand Education.

Sudo, K., Sekine, S., and Grishman, R. (2001). Automatic Pattern Acquisition for Japanese Information Extraction. In *Proc. of HLT*, San Diego, California.

Sukkarieh, J.Z., Pulman, S.G., and 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 (IAEA)*, Manchester, UK.

Sukkarieh, J.Z. and Pulman, S.G. (2005). Information extraction and machine learning: Auto-marking short free text responses to science questions. In *Proc. of AIED*.

Surdeanu, M., Harabagiu, S., Williams, J., and Aarseth, P. (2003). Using Predicate-Argument Structures for Information Extraction. *Proceedings of ACL-03*.

Sweet, A.P. and Snow, C.E. (Eds.). (2003). *Rethinking reading comprehension*. New York: Guilford Press.

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

Topping, K., and Whitley, M. (1990). Participant evaluation of parent-tutored and peer-tutored projects in reading. In *Educational Research, 32*(1), 14-32.

Torgesen, J.K., Wagner, R.K., and Rashotte, C. (1999). *Test of Word Reading Efficiency:TOWRE*. Austin, TX: PRO-ED.

Turney, P.D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, 491–502.

Turney, P.D., Littman, M.L., Bigham, J., and Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. *Proceedings of RANLP*, 482-489. Borovets, Bulgaria.

Vanderwende, L., Coughlin, D., and Dolan, B. (2005). What Syntax can Contribute in the Entailment Task. In *Proceedings of Pascal Challenge Workshop on Recognizing Textual Entailment*.

VanLehn, K., Lynch, C., Schulze, K. Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., and Wintersgill, M. (2005). The Andes physics tutoring system: Five years of evaluations. In G. McCalla and C. K. Looi (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence in Education*. Amsterdam: IOS Press.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., and Baggett, W.B. (2003). Why Do Only Some Events Cause Learning During Human Tutoring? In *Cognition and Instruction*, 21(3), 209–249. Lawrence Erlbaum Associates, Inc.

Vapnik, V. (1998). *Statistical learning theory*. New York: John Wiley and Sons Inc.

Ward, W.H. (1991) The Phoenix system: Understanding spontaneous speech. In *Proc. of IEEE ICASSP*.

Whittington, D., and Hunt, H. (1999). Approaches to the Computerised Assessment of Free-Text Responses. *Third International Computer Assisted Assessment Conference* Loughborough University June 1999.

Witten, I.H. and Frank E. (2000). *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of ACL-95*.

Zaenen, A., Karttunen, L., and Crouch, R. (2005). Local Textual Inference: can it be defined or circumscribed? In *Proc. of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, 31–36.