Learning to Assess Low-level Conceptual Understanding

Rodney D. Nielsen, Wayne Ward, and James H. Martin

Center For Computational Language and Education Research, Institute of Cognitive Science, University of Colorado
Campus Box 594, Boulder, CO 80309-0594
Rodney.Nielsen, Wayne.Ward, James.Martin @ Colorado.edu

Abstract

This paper analyzes the impact of several lexical and grammatical features in automated assessment of students' fine-grained understanding of tutored concepts. Truly effective dialog and pedagogy in Intelligent Tutoring Systems is only achievable when systems are able to understand the detailed relationships between a student's answer and the desired conceptual understanding. We describe a new method for recognizing whether a student's response entails that they understand the concepts being taught. We discuss the need for a finer-grained analysis of answers and describe a new representation for reference answers that addresses these issues, breaking them into detailed facets and annotating their relationships to the student's answer more precisely. Human annotation at this detailed level still results in substantial inter-annotator agreement, 86.0% with a Kappa statistic of 0.724. We present our approach to automatically assess student answers, which involves training machine learning classifiers on features extracted from dependency parses of the reference answer and student's response and features derived from domain-independent linguistic statistics. Our system's performance, 75.5% accuracy within domain and 65.9% out of domain, is encouraging and confirms the feasibility of the approach. Another significant contribution of this work is that the semantic assessment of answers is domain independent.

Introduction

The long-term goal of this work is to develop a domain-independent intelligent tutoring system (ITS) that can carry on a natural unrestricted dialog with a student and approach the two-sigma learning gains reported by Bloom for one-on-one tutoring (Bloom 1984). This paper focuses on two issues involved in the assessment of students' understanding necessary to achieve this goal. First, truly effective interaction and pedagogy is only possible if the automated tutor can assess student understanding at a relatively fine level of detail (c.f. Jordan, Makatchev, and VanLehn 2004). Second, natural unrestricted dialog requires domain-independent assessment of the learner's responses.

Still, most ITSs today provide only a shallow assessment of the learner's comprehension (e.g., a correct versus incorrect decision). Many ITS researchers are striving to provide more refined learner feedback (Graesser et al. 2001; Jordan, Makatchev, and VanLehn 2004; Peters et al. 2004; Roll et al. 2005; VanLehn et al. 2005); however, they are largely developing very domain-dependent approaches, requiring a significant investment in hand-crafted logic representations, parsers, knowledge-based ontologies, and or dialog control mechanisms. Similarly, research in the area of scoring free-text responses to short answer questions (e.g., Callear, Jerrams-Smith, and Soh 2001; Leacock 2004; Mitchell et al. 2002; Pulman and Sukkarieh 2005) also relies heavily on hand-crafted pattern rules, rather than being designed with the goal of accommodating dynamically generated, previously unseen questions, and that work does not provide feedback regarding the specific aspects of answers that are correct or incorrect. The PASCAL Recognizing Textual Entailment (RTE) challenge (Dagan, Glickman, and Magnini 2005) is addressing the task of domain-independent inference, but the task only requires systems to make yes-no judgments as to whether a human reading one text snippet would typically consider a second text to most likely be true.

This paper discusses a representation that facilitates answer assessment at a fine-grained level. We describe a corpus of elementary students' science assessments annotated with these fine-grained relationships. We present a domain-independent machine learning approach to automatically classify learner answers according to this scheme and report on our results. In the discussion, we detail the significance of the features we use to learn the classifiers and how they can be improved.

Annotating a Corpus

Fine-Grained Representation

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, we focus strictly on drawing inferences regarding the student's understanding of the low level concepts and relationships or facets of the reference answer. We use the word facet throughout this paper to generically refer to some part of a text's (or utterance's) meaning. The most common type of answer facet discussed is the meaning 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, we instead break the reference answer down into what we consider to be its lowest level compositional facets. This roughly translates to the set of triples composed of labeled dependencies in a dependency parse of the reference answer.[1] The following illustrates how a simple reference answer (1) is decomposed into the answer facets (1a-d) derived from its dependency parse. The dependency parse tree is shown in Figure 1. As can be seen in 1b and 1c, the dependencies are augmented by thematic roles (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. 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.

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

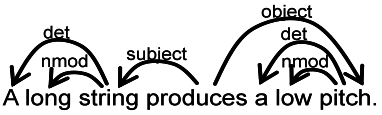

Figure 1. Dependency parse tree for example reference answer

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? We also need to break the annotation labels into finer levels in order to specify more clearly these relationships. These two issues – breaking the reference answer into fine grained facets and utilizing more expressive annotation labels – are an emphasis of the present work.

The Corpus

Our work focuses on the critical years of learning to comprehend text, K-6. We acquired data gathered from 3rd-6th grade students utilizing the Full Option Science System (FOSS), a proven system that has been in use across the country for over a decade. The data was gathered via the Assessing Science Knowledge (ASK) project, a research effort aligned with the FOSS program (Lawrence Hall of Science 2005).

[1] 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, usually labeled, between a governor (head) and a subordinate (modifier), and each node has a single governor (c.f., Nivre and Scholz 2004).

We used 290 free response questions taken from the ASK summative assessments covering the 16 diverse FOSS science teaching and learning modules (e.g., Mixtures, Landforms, Magnetism, etc). These questions had expected responses ranging from moderately short verb phrases to several sentences and could be assessed objectively; an Earth Materials question (Q) follows with its reference answer (R) and an example student answer (A).

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.

We generated a corpus from a random sample of the students' handwritten responses to these questions. While this corpus was not extracted from tutoring sessions, the questions and student answers are representative of the tutor we are currently developing based on the FOSS modules. The only special transcription instructions were to fix spelling errors (since these would be irrelevant in a spoken dialog environment, the target of our 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. We transcribed approximately 15,400 student responses (roughly 100 per question for three modules – Environment, Human Body and Water – to be used as a domain-independent test set for our algorithms and 40 per question for the remaining 13 modules). This resulted in about 144,000 total facet annotations.

The Annotation

Each reference answer (as specified by the ASK research team) was decomposed by hand into its constituent facets. Then for each student answer, the facets in the reference answer were annotated to describe whether and how they were addressed by the student. The reference answer facets, which are roughly extracted from the relations in a syntactic dependency parse and a shallow semantic parse, are modified slightly to either eliminate most function words or incorporate them into the relation labels (c.f., Lin and Pantel 2001). Example 2 illustrates the decomposition of one of the reference answers into its constituent parts along with their glosses.

(2) The string is tighter, so the pitch is higher.
(2a) Be(string, tighter)
(2a') The string is tighter.
(2b) Be(pitch, higher)
(2b') The pitch is higher.
(2c) Cause(2b, 2a)
(2c') 2b is caused by 2a

Table 1 presents the level of annotation labels expected to drive the tutor's dialog; see (Nielsen et al. 2008) for a description of the full set of annotation labels and further details on the annotation project.

Label	Brief Definition
Understood	Facets directly expressed or whose understanding is inferred
Contradicted	Facets contradicted by negation, antonyms, pragmatics, etc.
Self-Contra	Facets both contradicted and implied (self contradictions)
Diff-Arg	The core relation is expressed, but it has a different modifier
Unaddressed	Facets that are not addressed at all by the student's answer

Table 1. Facet Annotation Labels

We evaluated inter-annotator agreement on the 11 science modules double annotated at the time of writing, totaling 96,815 total facet annotations. Agreement was 86.0%, with a Kappa statistic of 0.724 corresponding with substantial agreement (Cohen 1960).

Automated Classification

A high level description of the system classification procedure is as follows. We start with hand generated reference answer facets, similar to typed dependency triples. We generate automatic parses for the reference answers and the student answers. Then for each student answer, we generate a training (or test) example for each facet of the associated reference answer. These examples are comprised of features extracted from the answers and their dependency parses, as described in Table 2. Finally, we train a machine learning classifier on our training data and use it to classify the unseen examples in the test set according to the labels in Table 1.

Preprocessing and Representation

Many of the machine learning features described here are based on document co-occurrence counts. Rather than use the web to extract these statistics (as did Turney (2001) and Glickman et al. (2005), who generate comparable metrics), we use three publicly available corpora (Gigaword, Reuters, and Tipster) totaling 7.4M articles and 2.6B indexed terms. These corpora are all drawn from the news domain, making them less than ideal for assessing student answers to science questions, but this does highlight the domain-independent nature of our approach. The reason we used these corpora to extract statistics indicating the similarity of words in our student answers to those in the reference answers was simply because they were readily available and indexed. In future work, we will utilize corpora believed to be more appropriate for the task. This should easily result in an improvement in assessment performance, particularly since much of our science test data includes vocabulary that is not used in the newswire text. The above corpora were indexed and searched using Lucene, a publicly available Information Retrieval tool. We used their PorterStemFilter to replace the surface form of words with their lexical stem.

Before extracting features, we generate dependency parses of the reference answers and student answers using MaltParser (Nivre et al. 2007). These parses are automatically modified by reattaching auxiliary verbs and their modifiers to the associated regular verbs, incorporating prepositions into the dependency relation labels (c.f. Lin and Pantel 2001) and similarly appending negation terms onto the associated dependency relations. These modifications increase the likelihood that terms carrying significant semantic content are joined by dependencies that will be the focus of feature extraction.

Machine Learning Features

We investigated a variety of linguistic features and chose to analyze the features summarized in Table 2, 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, due to space limitations we focus on this set of features in our results and the feature analysis section which follows. The features assess lexical similarity via part of speech (POS) tags, lexical stem matches, and lexical entailment probabilities following (Glickman et al. 2005). They include dependency parse information such as relevant dependency relation types and path edit distances. Remaining features include information about polarity among other indicators.

Lexical Features

Gov/Mod_MLE: The lexical entailment probabilities (LEPs) for the reference answer (RA) facet governor (Gov; eg, *produce* in 1b) and modifier (Mod; eg, *string* in 1b) following (Glickman et al. 2005; c.f., Turney 2001). {2a: the LEPs for *string→string* and *tension→tighter*, respectively}*

Gov/Mod_Match: True if the Gov (Mod) stem has an exact match in the learner answer. {2a: True and (False), respectively}*

Subordinate_MLEs: The LEPs for the primary constituent facets' Govs and Mods when the facet represents a relation between propositions. {2c: the LEPs for *pitch→pitch*, *up→higher*, *string→string*, and *tension→tighter*}*

Syntactic Features

Gov/Mod_POS: Part of speech tags for the RA facet Gov & Mod

Facet/AlignedDep_Reltn: The dependency (Dep) labels of the facet and aligned learner answer (LA) Dep – Dep alignments were based on co-occurrence MLEs as with words, i.e., they estimate the likelihood of seeing the RA Dep in a document given it contains the LA Dep. {2a: Be and Have, respectively}*

Dep_Path_Edit_Dist: The edit distance between the Dep path connecting the facet's Gov and Mod (not necessarily a single step due to parser errors) and the path connecting the aligned terms in the LA. Paths include the Dep relations with their attached prepositions, negations, etc, the direction of each Dep, and the POS tags of the terms on the path. The calculation applied 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). {2b: *Distance*(*up*:VMod>*went*:V<*pitch*:Subject, *pitch*:Be>*higher*)}*

Other Features

Consistent_Negation: True if the RA facet and aligned LA dependency path had a single negation or if neither had a negation.

RA_CW_cnt: Number of content words in the RA (based on Lucene's stop-word list). {5 (*string, tighter, so, pitch* & *higher*)}*

*Examples within {} are based on RA Example 2 and the learner answer: *The pitch went up because the string has more tension*

Table 2. Machine Learning Features

Machine Learning Approach

The 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 16 science modules (Environment, Human Body and Water), providing what is considered 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 22 randomly selected questions from the 233 questions in the remaining 13 modules, providing a question-independent test set from within the same domain or topic areas seen in the training data. 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 had been adjudicated; whereas, about half of the remaining data (training data, Unseen Questions and Unseen Answers) had not. We used the most recent annotation of the unadjudicated data during the experiments presented here. Facets assumed to be understood a priori, generally because the information was given in the question, are tagged as such in the reference answer markup. For proper classification these facets require special logic and features that we have not yet implemented, so all *Assumed* facets were withheld from the datasets (*Assumed* facets comprise 33% of the data). We will address this in future work, since in a question- or domain-independent assessment, it is necessary to identify what information should be assumed to be understood a priori. This selection resulted in a total of 54,967 training examples, 30,514 examples in the Unseen Modules test set, 6,699 in the Unseen Questions test set and 3,159 examples in the Unseen Answers test set.

We evaluated several machine learning algorithms and C4.5 (Quinlan 1993) achieved the best results in cross validation on the training data. Therefore, we used it to obtain all of the results presented here for this new task of automatically annotating low-level reference answer facets with fine-grained classifications.[2]

System Results and Analysis

Table 3 shows the classifier's accuracy in cross-validation on the training set as well as on each of our test sets. The first two columns provide baselines: 1) the accuracy of a classifier that always outputs the label of the most frequent class in the training set (Unaddressed) and 2) the accuracy based on a lexical decision that chooses Understood, when both the facet's governing term and its modifier are present in the learner's answer, and outputs Unaddressed otherwise. (We also tried placing a threshold on the product of the lexical entailment probabilities of the facet's governor and modifier, following Glickman et al. (2005), who achieved the best results in the first RTE challenge, but this gave virtually the same results as the word matching baseline). The column labeled Table 2 Features presents the results of our classifier. (Reduced Training is described in the Discussion and Future Work section below.)

	Majority Label	Lexical Baseline	Table 2 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 3. Classifier Accuracy

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 label in the training set) and relative to the 77.05% accuracy of a classifier built using all of the features in Table 2. Positive values in the second column indicate that the feature hurt the classifier's ability to generalize to held-out cross-validation data. 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 (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 and seem not to add value to the final classifier.

Feature Added or Removed	**Change from 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 4. Feature Impact relative to 1) Baseline, 2) All Features

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

[2] Many other classifiers performed comparably and Random Forests outperformed C4.5 in previous evaluations. A thorough analysis of the impact of the classifier chosen has not been completed at this time.

the baseline, but in combination with other features fail to improve the final system's performance. This implies that we must perform a thorough error analysis of this subset of facets and design an alternate approach to 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 dependency path edit distance feature demonstrates that deeper syntax does help assess understanding at this fine-grained level. Others have also found syntactic analysis to improve over a purely lexical approach (e.g., Jordan, Makatchev and VanLehn 2004). 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.

Discussion and Future Work

Achieving results similar to those for the Unseen Answers test set is dependent on time-consuming hand-annotation of roughly 36 answers for each new question. Whereas, no additional data collection, annotation, or system training should be required to achieve the results associated with our Unseen Modules test set. Here, the data used for training the classifier did not include any answers to questions in the test domain – the training data is from entirely different fields of science. However, to consider this to be a truly domain-independent approach, we must move away from hand-generated reference answer facets by adapting existing automated parsers to children's dialogue and our task.

This is a novel task and a new dataset, and the results relative to the most frequent class label baseline are promising. The accuracy is 24.4%, 3.3%, and 8.0% over this baseline 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. It seemed probable that the decision tree 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 our Unseen Answers test set, but would negatively affect performance on less similar data, such as our Unseen Questions and Modules test sets. To test this hypothesis, we reduced the number of examples in our training set to about 8,000, which would result in fewer of these dense clusters, and retrained the classifier. The results, shown in the Reduced Training column, were improvements of 4.8% and 4.5% on the Unseen Questions and Modules test sets, respectively. We must now 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.

We are performing a more detailed feature and error analysis currently, but Table 4 shows that the most salient features are simple lexical features (e.g., lexical co-occurrence statistics). The simple lexical baseline in Table 3 shows an average improvement of about 6.5% relative to classifying according to the most frequent class label in the training set. We believe two of the biggest sources of error derive from the lack of pronoun resolution and errors propagated from POS tags and dependency parses. 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. Analyzing the dependency parses of 51 of the student answers, many have what we believe 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). We are working on approaches to mitigate this problem.

In prior work, we achieved confidence weighted scores approximately 10% (absolute) over the classification accuracy, indicating that the class probability estimates will be useful to the dialog manager in deciding how strongly to believe in the classifier's output. 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. These results demonstrate that even if there is a moderate error rate in classifying the facets, low confidence estimates can still result in effective dialogue. We are currently working on integrating this assessment technology into our science tutor, which will cover several of the FOSS elementary school science modules.

Conclusion

The results presented here for the Unseen Answers test set are 19.4% over a lexical baseline derived from the best system at the first RTE challenge. The out-of-domain results are 12.5% and 3% over the most frequent class and lexical baselines respectively. This demonstrates that the task is feasible and we are currently improving the feature set and integrating the system with a science tutor to evaluate its efficacy. Even when the prediction is not correct as long as the tutor acts according to the confidence, the dialog can be effective. Two of the most significant contributions of this work are defining and evaluating a 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 in the future, and laying the framework for an answer assessment system that can classify learner responses to previously unseen questions according to this scheme.

The corpus of learner answers described here was annotated with substantial agreement (86.0%, Kappa = 0.724) and will be made publicly available for other researchers to utilize in improving their tutoring and educational assessment technologies. 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. To our knowledge, this is also the first work to demonstrate success in assessing roughly sentence length constructed responses from elementary school children.

Prior work on intelligent tutoring systems has largely 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.

Acknowledgements

We are grateful to the anonymous reviewers, whose comments were extremely detailed and useful. This work was partially funded by Award Number 0551723 of the NSF.

References

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

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

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

Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*

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

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

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

Jordan PW, Makatchev M, VanLehn K (2004) Combining competing language understanding approaches in an intelligent tutoring system. In Lester JC, Vicari RM, Paraguacu F (eds) *7th Conference on Intelligent Tutoring Systems*, 346-357. Springer-Verlag, Berlin Heidelberg New York

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

Kohavi R (1995) The power of decision tables. In *Proc. of the eighth European Conference on Machine Learning*

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)

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

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

Nielsen RD, Ward W, Martin J, Palmer M (2008) Annotating kids' understanding of science concepts. In *Proc LREC*

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

Nivre, J. and Scholz, M. (2004). Deterministic dependency parsing of English text. In *Proc COLING*.

Peters S, Bratt EO, Clark B, Pon-Barry H, Schultz K (2004) Intelligent systems for training damage control assistants. In *Proc. of Interservice/Industry Training, Simulation and Education Conference*

Pulman SG, Sukkarieh JZ (2005) Automatic short answer marking. In *Proc. of the 2nd Workshop on Building Educational Applications Using NLP, ACL*

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

Roll I, Baker RS, Aleven V, McLaren BM, Koedinger KR (2005) Modeling students' metacognitive errors in two intelligent tutoring systems. In Ardissono L, Brna P, Mitrovic A (eds) *User Modeling*, pp 379–388

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

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

VanLehn K, Lynch C, Schulze K Shapiro JA, Shelby R, Taylor L, Treacy D, Weinstein A, Wintersgill M (2005) The Andes physics tutoring system: Five years of evaluations. In McCalla G, Looi CK (eds) *Proc. of the 12th International Conference on AI in Education*