Classification Errors in a Domain-Independent Assessment System

Rodney D. Nielsen[1,2], Wayne Ward[1,2] and James H. Martin[1]
[1] Center for Computational Language and Education Research, University of Colorado, Boulder
[2] Boulder Language Technologies, 2960 Center Green Ct., Boulder, CO 80301
Rodney.Nielsen, Wayne.Ward, James.Martin@Colorado.edu

Abstract

We present a domain-independent technique for assessing learners' constructed responses. The system exceeds the accuracy of the majority class baseline by 15.4% and a lexical baseline by 5.9%. The emphasis of this paper is to provide an error analysis of performance, describing the types of errors committed, their frequency, and some issues in their resolution.

1 Introduction

Assessment within state of the art Intelligent Tutoring Systems (ITSs) generally provides little more than an indication that the student's response expressed the target knowledge or it did not. There is no indication of exactly what facets of the concept a student contradicted or failed to express. Furthermore, virtually all ITSs are developed in a very domain-specific way, with each new question requiring the handcrafting of new semantic extraction frames, parsers, logic representations, or knowledge-based ontologies (c.f., Graesser et al., 2001; Jordan et al., 2004; Peters et al., 2004; Roll et al., 2005; VanLehn et al., 2005). This is also true of research in the area of scoring constructed response questions (e.g., Callear et al., 2001; Leacock, 2004; Mitchell et al., 2002; Pulman and Sukkarieh, 2005). The present paper analyzes the errors of a system that was designed to address these limitations.

Rather than have a single expressed versus not-expressed assessment of the reference answer as a whole, we instead break the reference answer down into what we 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. Breaking the reference answer down into fine-grained facets permits a more focused 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, (e.g., did the student contradict the facet or just fail to address it?). Therefore, 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 facet.

In this paper, we present an error analysis of our system, detailing the most frequent types of errors encountered in our implementation of a domain-independent ITS assessment component and discuss plans for correcting or mitigating some of the errors. The system expects constructed responses of a phrase to a few sentences, but does not rely on technology developed specifically for the domain or subject matter being tutored – without changes, it should handle history as easily as science. We first briefly describe the corpus used, the knowledge representation, and the annotation. In section 3, we describe our assessment system. Then we present the error analysis and discussion.

2 Assessing Student Answers

2.1 Corpus

We acquired grade 3-6 responses to 287 questions from the Assessing Science Knowledge (ASK) project (Lawrence Hall of Science, 2006). The responses, which range in length from moderately

Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications, pages 10–18,
Columbus, Ohio, USA, June 2008.

short verb phrases to several sentences, cover all 16 diverse teaching and learning modules, spanning life science, physical science, earth and space science, scientific reasoning, and technology. We generated a corpus by transcribing a random sample (approx. 15400) of the students' handwritten responses.

2.2 Knowledge Representation

The ASK assessments included a reference answer for each of their constructed response questions. We decomposed these reference answers into low-level facets, roughly extracted from the relations in a syntactic dependency parse and a shallow semantic parse. However, we use the word *facet* to refer to any fine-grained component of the reference answer semantics. The decomposition is based closely on these well-established frameworks, since the representations have been shown to be learnable by automatic systems (c.f., Gildea and Jurafsky, 2002; Nivre et al., 2006). These facets are the basis for assessing learner answers. See (Nielsen et al., 2008b) for details on extracting the facets; here we simply sketch the makeup of the final assessed reference answer facets.

Example 1 presents a reference answer from the Magnetism and Electricity module and illustrates the facets derived from its dependency parse (shown in Figure 1), along with their glosses. These facets represent the fine-grained knowledge the student is expected to address in their response.

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

(1a) NMod(ring, brass)
(1a') The ring is brass.
(1b) Theme_not(stick, ring)
(1b') The ring does not stick.
(1c) Destination_to_not(stick, nail)
(1c') Something does not stick to the nail.
(1d) Be_not(ring, iron)
(1d') The ring is not iron.
(1e) Cause_because(1b-c, 1d)
(1e') 1b and 1c are caused by 1d.

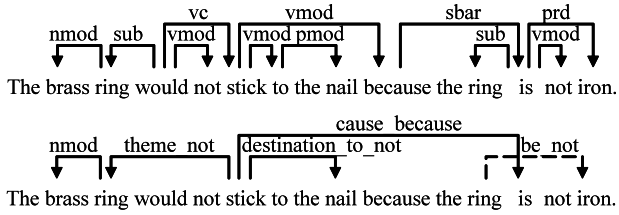

Figure 1. Reference answer representation revisions

Typical facets, as in (1a), are derived directly from a dependency parse, in this case retaining its dependency type label, NMod (noun modifier). Other facets, such as (1b-e), are the result of combining multiple dependencies, VMod(*stick*, *to*) and PMod(*to*, *nail*) in the case of (1c). When the head of the dependency is a verb, as in (1b,c), we use Thematic Roles from VerbNet (Kipper et al., 2000) and adjuncts from PropBank (Palmer et al., 2005) to label the facet relation. Some copulas and similar verbs were themselves used as facet relations, as in (1d). Dependencies involving determiners and many modals, such as *would*, in ex. 1, are discarded and negations, such as *not*, are incorporated into the associated facets.

We refer to facets that express relations between higher-level propositions as *inter-propositional facets*. An example of such a facet is (1e) 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 1e), we also note up to two key facets from each of the propositions that the relation is connecting (b, c, and d in ex. 1). Reference answer facets that are assumed to be understood by the learner a priori, (generally because they are part of the information given in the question), are also annotated to indicate this.

There were a total of 2878 reference answer facets, with a mean of 10 facets per reference answer (median of 8). 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 (section 3) focus on the facets that are not assumed to be understood a priori (67% of all facets); of these, 12% are inter-propositional.

2.3 Annotating Student Understanding

After defining the reference answer facets, we annotated each student answer to indicate whether and how they addressed each reference answer facet. We settled on the annotation labels in Table 1. For a given student answer, one label is assigned for each facet in the associated reference answer. These labels and the annotation process are detailed in (Nielsen et al., 2008a).

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

Table 1. Facet Annotation Labels

Example 2 shows a fragment of a question and associated 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 (3) along with its final annotation in 2a'-c'. It is assumed that the student understands that the pitch is higher (facet 2b), since this is given in the question and similarly it is assumed that the student will be explaining what has the causal effect of producing this higher pitch (facet 2c). Therefore, unless the student explicitly addresses these facets they are labeled *Assumed*. The student phrase *the string is long* is aligned with reference answer facet 2a, since they are both expressing a property of the string, but since the phrase neither contradicts nor indicates an understanding of the facet, the facet is labeled *Diff-Arg*, 2a'. The causal facet 2c' is labeled *Expressed*, since the student expresses a causal relation and the cause and effect are each properly aligned. In this way, the automated tutor will know the student is on track in attempting to address the cause and it can focus on remediating the student's understanding of that cause.

(2) Question: ... 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...

(2a) Be(string, tighter), ---
(2b) Be(pitch, higher), Assumed
(2c) Cause(2b, 2a), Assumed

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

(2a') Be(string, tighter), Diff-Arg
(2b') Be(pitch, higher), Expressed
(2c') Cause(2b', 2a'), Expressed

A tutor will treat the labels Expressed, Inferred and Assumed all as Understood by the student and similarly Contra-Expr and Contra-Infr are combined as Contradicted. These labels are kept separate in the annotation to facilitate training different systems to detect these different inference relationships, as well as to allow evaluation at that level. The consolidated set of labels, comprised of Understood, Contradicted, Self-Contra, Diff-Arg and Unaddressed, are referred to as the *Tutor Labels*.

3 Automated Classification

A high level description of the assessment procedure is as follows. We start with the hand generated reference answer facets. We generate automatic parses for the reference answers and the student answers and automatically modify these parses to match our desired representation. Then for each reference answer facet, we extract features indicative of the student's understanding of that facet. Finally, we train a machine learning classifier on our training data and use it to classify unseen test examples, assigning a Tutor Label (described in the preceding paragraph), for each reference answer facet.

3.1 Preprocessing and Representation

Many of the features utilized by the machine learning algorithm here are based on document co-occurrence counts. We use three publicly available corpora (English Gigaword, The Reuters corpus, and Tipster) totaling 7.4M articles and 2.6B terms. These corpora are all drawn from the news domain, making them less than ideal sources for assessing student's answers to science questions. We utilized these corpora to generate term relatedness statistics primarily because they comprised a readily available large body of text. They were indexed and searched using Lucene, a publicly available Information Retrieval tool.

Before extracting features, we automatically generate dependency parses of the reference answers and student answers using MaltParser (Nivre

et al., 2006). These parses are then *automatically* modified in a way similar to the manual revisions made when extracting the reference answer facets, as sketched in section 2.2. We reattach auxiliary verbs and their modifiers to the associated regular verbs. We incorporate prepositions and copulas into the dependency relation labels, and similarly append negation terms onto the associated dependency relations. These modifications, all made automatically, increase the likelihood that terms carrying significant semantic content are joined by dependencies that are utilized in feature extraction. In the present work, we did not make use of a thematic role labeler.

3.2 Machine Learning Features & Approach

We investigated a variety of linguistic features and chose to utilize the features summarized in Table 2, informed by training set cross validation results. The features assess the facets' lexical similarity via lexical entailment probabilities following (Glickman et al., 2005), part of speech (POS) tags, and lexical stem matches. They include syntactic information extracted from the modified dependency parses such as relevant relation types and path edit distances. Remaining features include information about polarity among other things. The revised dependency parses described earlier are used in aligning the terms and facet-level information for feature extraction, as indicated in the feature descriptions.

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, providing a domain-independent test set. The second, *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. 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. In the present work, we utilize only the facets that were not 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.

Lexical Features
Gov/Mod_MLE: The lexical entailment probabilities (LEPs) for the reference answer facet governor (Gov; e.g., *string* in 2a) and modifier (Mod; e.g., *tighter* in 2a) following (Glickman et al., 2005; c.f., Turney, 2001). The LEP of a reference answer word *w* is defined as: (1) $LEP(w) = \max_{v \in s}\left(n_{w,v}/n_v\right)$, where *v* is a word in the student answer, n_v is the # of docs (see section 3.1) containing *v*, and $n_{w,v}$ is the # of docs where *w* & *v* cooccur. {Ex. 2a: the LEPs for *string→string* and *tension→ tighter*, respectively}†
Gov/Mod_Match: True if the Gov (Mod) stem has an exact match in learner answer. {Ex. 2a: True for Gov: *string*, and (False for Mod: no stem match for *tighter*)}†
Subordinate_MLEs: The lexical entailment probabilities for the primary constituent facets' Govs and Mods when the facet represents a relation between higher-level propositions (see inter-propositional facet definition in section 2.2). {Ex. 2c: the LEPs for *pitch→pitch*, *up→higher*, *string→string*, and *tension→tighter*}†
Syntactic Features
Gov/Mod_POS: POS tags for the facet's Gov and (Mod). {Ex. 2a: NN for *string* and (JJR for *tighter*)}†
Facet/AlignedDep_Reltn: The labels of the facet and aligned learner answer dependency – alignments were based on co-occurrence MLEs as with words, (i.e., they estimate the likelihood of seeing the reference answer dependency in a document given it contains the learner answer dependency – replace words with dependencies in equation 1 above). {Ex. 2a: Be is the facet label and Have is the aligned student answer dependency}†
Dep_Path_Edit_Dist: The edit distance between the dependency 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 learner answer. Paths include the dependency relations generated in our modified parse with their attached prepositions, negations, etc, the direction of each dependency, and the POS tags of the terms on the path. The calculation applies heuristics to judge the similarity of each part of the path (e.g., dropping a subject had a much higher cost than dropping an adjective). Alignment for this feature was made based on which set of terms in an *N*-best list (*N*=5 in the present experiments) for the Gov and Mod resulted in the smallest edit distance. The *N*-best list was generated based on the lexical entailment values (see Gov/Mod_MLE). {Ex. 2b: *Distance*(*up*:VMod> *went*:V<*pitch*:Subject, *pitch*:Be>*higher*)}†
Other Features
Consistent_Negation: True if the facet and aligned student dependency path had the same number of negations. {Ex. 2a: True: neither one have a negation}†
RA_CW_cnt: The number of content words (non-function words) in the reference answer. {Ex. 2: 5 = count(*string, tighter, so, pitch & higher*)}†
† Examples within {} braces are based on reference answer Ex. 2 and the learner answer: *The pitch went up because the string has more tension*

Table 2. Machine Learning Features

We evaluated several machine learning algorithms (rules, trees, boosting, ensembles and an svm) 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. A number of classifiers performed comparably and Random Forests outperformed C4.5 with a previous feature set and subset of data. A thorough analysis of the impact of the classifier chosen has not been completed at this time.

3.3 System Results

Given a student answer, we generate a separate Tutor Label (described at the end of section 2.3) for each associated reference answer facet to indicate the level of understanding expressed in the student's answer (similar to giving multiple marks on a test). Table 3 shows the classifier's Tutor Label accuracy over all reference answer facets in cross validation on the training set as well as on each of our test sets. The columns first show two simpler baselines, the accuracy of a classifier that always chooses the most frequent 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, (we also tried placing a threshold on the product of the governor and modifier lexical entailment probabilities following Glickman et al. (2005), who achieved the best results in the first RTE challenge, but this gave virtually the same results as the word matching baseline). The column labeled Table 2 Features presents the results of our classifier. (Reduced Training is described in the Discussion section, which follows.)

	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	**68.8**

Table 3. Classifier Accuracy

4 Discussion and Error Analysis

4.1 Results Discussion

The accuracy achieved, assessing learner answers within this new representation framework, represent an improvement of 24.4%, 3.3%, and 8.0% over the majority class 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 other two 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.

To test this hypothesis, we reduced the size of our training set to about 8,000 randomly selected examples, which would result in fewer of these dense clusters, and retrained the classifier. The result for Unseen Questions, shown in the *Reduced Training* column, was an improvement of 4.8%. Given this promising improvement, we attempted to find the optimal training set size through cross-validation on the training data. Specifically, we iterated over the science modules holding one module out, training on the other 12 and testing on the held out module. We analyzed the learning curve varying the number of randomly selected examples per facet. We found the optimal accuracy for training set cross-validation by averaging the results over all the modules and then trained a classifier on that number of random examples per facet in the training set and tested on the Unseen Modules test set. The result was an increase in accuracy of 7.4% over training on the full training set. In future work, we will investigate other more principled techniques to avoid this type of overfitting, which we believe is somewhat atypical.

4.2 Error Analysis

In order to focus future work on the areas most likely to benefit the system, an error analysis was performed based on the results of 13-fold cross-validation on the training data (one fold per science module). In other words, 13 C4.5 decision tree classifiers were built, one for each science module in the training set; each classifier was trained,

utilizing the feature set shown in Table 2, on all of the data from 12 science modules and then tested on the data in the remaining, held-out module. This effectively simulates the Unseen Modules test condition. To our knowledge, no prior work has analyzed the assessment errors of such a domain-independent ITS.

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. 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 almost 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. We discuss Expressed facets in the next section of the paper and Unaddressed in the following section.

4.3 Errors in Expressed Facets

Without examining each example relative to the decision tree that classified it, it is not possible to know exactly what caused the errors. The analysis here simply indicates what factors are involved in inferring whether the reference answer facets were understood and what relationships exist between the student answer and the reference answer facet. We analyzed 100 random examples of errors where annotators considered the facet Expressed and the system labeled it Unaddressed, but the analysis only considered one example for any given reference answer facet. Out of these 100 examples, only one looked as if it was probably incorrectly annotated. We group the potential error factors seen in the data, listed in order of frequency, according to issues associated with paraphrases, logical inference, pragmatics, and preprocessing errors. In the following paragraphs, these groups are broken down for a more fine-grained analysis. In over half of the errors considered, there were two or more of these fine-grained factors involved.

Paraphrase issues, taken broadly, are subdivided into three main categories: coreference resolution, lexical substitution, syntactic alternation and phrase-based paraphrases. Our results in this area are in line with (Bar-Haim et al., 2005), who considered which inference factors are involved in proving textual entailment. 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 single alignment that gives the learner the most credit. In 6 errors, the student referred to a concept by another term (e.g., substituting *stuff* for *pieces*). In another 6 errors, the student used one of the terms in a noun phrase from either the question or reference answer 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 (one could argue that there is an important distinction between the answers, but requiring elementary school students to answer at this level of specificity could result in an overwhelming number of interactions to clarify understanding).

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 broad coverage lexical resources. For example, substituting *tiny* for *small*, CO_2 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*. VerbNet also does not show a relation between either pair of words. Concept definitions account for an additional 14 issues that could potentially be addressed by lexical resources such as WordNet.

Vanderwende et al. (2005) found that 34% of the Recognizing Textual Entailment Challenge test data could be handled by recognizing simple syntactic variations. However, while syntactic variation is certainly common in the kids' data, it did not appear to be the primary factor in any of the system errors. Most of the remaining paraphrase errors were classified as involving phrase-based paraphrases. Examples here include *...it will heat up faster* versus *it got hotter faster* and *in the middle* versus *halfway between*. Six related 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*). Paraphrase recognition is an area that we intend to invest significant time in future research (c.f., Lin and Pantel, 2001; Dolan et al., 2004). This research should also reduce the error rate on lexical paraphrases.

The next most common issues after paraphrases were deep or logical reasoning and then pragmatics. These two factors were involved in nearly 40% of the errors. Examples of logical inference include recognizing that two 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. It is interesting that these are all examples that three annotators unanimously considered to be Expressed versus Inferred facets.

Finally, the remaining errors were largely the result of preprocessing issues. At least two errors would be eliminated by simple data normalization (*3→three* and *g→grams)*. 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 labels 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.

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*. 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 sentences when two noun phrases are conjoined and they occur between two verbs, as is the case in the preceding example, where the alternative segmentation results in correct parses. Then the system could choose the parse that is most consistent with the reference answer. While we believe improving the parser output will result in higher accuracy by the assessment classifier, there was little evidence to support this in the small number of parses examined in the assessment error analysis. We only checked the parses when the dependency path features looked wrong and it was somewhat surprising that the classifier made an error (for example, when there were simple lexical substitutions involving very similar words) – this was the case for only about 10-15 examples. Only two of these classification errors were associated with parser errors. However, better parses should lead to more reliable (less noisy) features, which in turn will allow the machine learning algorithm to more easily recognize which features are the most predictive.

It should be emphasized that over half of the errors in Expressed facets involved more than one

of the fine-grained factors discussed here. For example, to recognize the child understands a tree is blocking the sunlight based on the answer *There is a shadow there because the sun is behind it and light cannot go through solid objects. 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 recognizing that *light cannot go through* [*the tree*] entails the tree blocks the light.

4.4 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*.

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 the production of steel, 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.

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

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

5 Conclusion

To our knowledge, this is the first work to successfully assess constructed-response answers from elementary school students. We achieved promising results, 24.4% and 15.4% over the majority class baselines for Unseen Answers and Unseen Modules, respectively. The annotated corpus associated with this work will be made available as a public resource for other researches working on educational assessment applications or other textual entailment applications.

The focus of this paper was to provide an error analysis of the domain-independent (Unseen Modules) assessment condition. We discussed the common types of issues involved in errors and their frequency when assessing young students' understanding of the fine-grained facets of reference answers. This domain-independent assessment will facilitate quicker adaptation of tutoring systems (or general test assessment systems) to new topics, avoiding the need for a significant effort in hand-crafting new system components. It is also a necessary prerequisite to enabling unrestricted dialogue in tutoring systems.

Acknowledgements

We would like to thank the anonymous reviewers, whose comments improved the final paper. This work was partially funded by Award Number 0551723 from the National Science Foundation.

References

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

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

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

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

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

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.

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.

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.

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

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.

Nielsen, R., Ward, W., Martin, J. and Palmer, M. 2008a. Annotating Students' Understanding of Science Concepts. In *Proc. LREC*.

Nielsen, R., Ward, W., Martin, J. and Palmer, M. 2008b. Extracting a Representation from Text for Semantic Analysis. In *Proc. ACL-HLT*.

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., 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)*.

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

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*.

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.

Roll, WE., 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.

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.

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

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.