# Extracting a Representation from Text for Semantic Analysis

**Rodney D. Nielsen[1,2], Wayne Ward[1,2], James H. Martin[1], and Martha Palmer[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, Martha.Palmer@Colorado.edu

## Abstract

We present a novel fine-grained semantic representation of text and an approach to constructing it. This representation is largely extractable by today's technologies and facilitates more detailed semantic analysis. We discuss the requirements driving the representation, suggest how it might be of value in the automated tutoring domain, and provide evidence of its validity.

## 1 Introduction

This paper presents a new semantic representation intended to allow more detailed assessment of student responses to questions from an intelligent tutoring system (ITS). Assessment within current ITSs generally provides little more than an indication that the student's response expressed the target knowledge or it did not. 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., Jordan et al., 2004). This is also true of research in the area of scoring constructed response questions (e.g., Leacock, 2004).

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

Rather than have a single expressed versus unexpressed 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 completely 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. The emphasis of this paper is on this fine-grained facet-based representation – considerations in defining it, the process of extracting it, and the benefit of using it.

## 2 Representing the Target Knowledge

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 short verb phrases to several sentences, cover all 16 diverse Full Option Science System 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.



*Proceedings of ACL-08: HLT, Short Papers (Companion Volume)*, pages 241–244,
Columbus, Ohio, USA, June 2008. 

## 2.1 Knowledge Representation

The ASK assessments included a reference answer for each constructed response question. These reference answers were manually decomposed into fine-grained facets, roughly extracted from the relations in a syntactic dependency parse and a shallow semantic parse. 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).

Figure 1 illustrates the process of deriving the constituent facets that comprise the representation of the final reference answer. We begin by determining the dependency parse following the style of MaltParser (Nivre et al., 2006). This dependency parse was then modified in several ways. The rationale for the modifications, which we elaborate below, is to increase the semantic content of facets. These more expressive facets are used later to generate features for the assessment classification task. These types of modifications to the parser output address known limitations of current statistical parser outputs, and are reminiscent of the modifications advocated by Briscoe and Carroll for more effective parser evaluation, (Briscoe, et. al, 2002). Example 1 illustrates the reference answer facets derived from the final dependencies in Figure 1, along with their glosses.

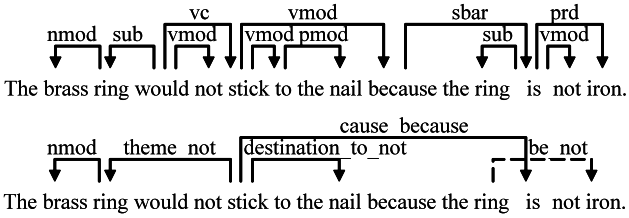


Figure 1. Reference answer representation revisions

(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.

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. In our example, the verb *stick* is made the governor of *would*, whose modifiers are reattached to *stick*. Similarly, the noun phrases *the pattern of pigments* and *the bunch of leaves* typically result in identical dependency parses. However, the word *pattern* is considered the governor of *pigments*; whereas, conversely the word *leaves* is treated as the governor of *bunch* because it carries more semantics. Then, terms that were not crucial to the student answer, frequently auxiliary verbs, were removed (e.g., the modal *would* and determiners in our example).

Next, we incorporate prepositions into the dependency type labels following (Lin and Pantel, 2001). This results in the two dependencies vmod(*stick, to*) and pmod(*to, nail*), each of which carries little semantic value over its key lexical item, *stick* and *nail*, being combined into the single, more expressive dependency vmod_to(*stick, nail*), ultimately vmod is replaced with destination, as described below. Likewise, the dependencies connected by *because* are consolidated and *because* is integrated into the new dependency type.

Next, copulas and a few similar verbs are also incorporated into the dependency types. The verb's predicate is reattached to its subject, which becomes the governor, and the dependency is labeled with the verb's root. In our example, the two semantically impoverished dependencies sub(*is, ring*) and prd(*is, iron*) are combined to form the more meaningful dependency be(*ring, iron*). Then terms of negation are similarly incorporated into the dependency types.

Finally, wherever a shallow semantic parse would identify a predicate argument structure, we used the thematic role labels in VerbNet (Kipper et al., 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. For example, in the sentence *As it freezes the water will expand and crack the glass*, typically the dependency between *crack* and its subject *water* is not generated since it would lead to a non-projective tree, but it does play the role of Agent in a semantic parse. In a small number of instances, these labels were also at-

tached 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*).

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 example 1). Reference answer facets that are assumed to be understood by the learner a priori, (e.g., because they are part of the question), are also annotated to indicate this.

There were a total of 2878 reference answer facets, resulting in a mean of 10 facets per answer (median 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 results of 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.

A total of 36 different facet relation types were utilized. The majority, 21, are VerbNet thematic roles. Direction, Manner, and Purpose are PropBank adjunctive argument labels (Palmer et al., 2005). Quantifier, Means, Cause-to-Know and copulas were added to the preceding roles. 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, etc., 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, Patient, AMod, and Location, which together account for 70% of the reference answer facet relations

### 2.2 Student Answer Annotation

For each student answer, we annotated each reference answer facet to indicate whether and how the student addressed that facet. We settled on the five annotation categories in Table 1. These labels and the annotation process are detailed in (Nielsen et al., 2008b).

| |
|---|
| **Understood**: Reference answer facets directly expressed or whose understanding is inferred |
| **Contradiction**: Reference answer facets contradicted by negation, antonymous expressions, pragmatics, etc. |
| **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

## 3 Automated Classification

As partial validation of this knowledge representation, we present results of an automatic assessment of our student answers. 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 training data and use it to classify unseen test examples, assigning a Table 1 label for each reference answer facet.

We used a variety of linguistic features that assess the facets' similarity via lexical entailment probabilities following (Glickman et al., 2005), part of speech tags and lexical stem matches. They include information extracted from modified dependency parses such as relevant relation types and path edit distances. Revised dependency parses are used to align the terms and facet-level information for feature extraction. Remaining details can be found in (Nielsen et al., 2008a) and are not central to the semantic representation focus of this paper. Current classification accuracy, assigning a Table 1 label to each reference answer facet to indicate the student's expressed understanding, is 79% within domain (assessing unseen answers to questions associated with the training data) and 69% out of domain (assessing answers to questions regarding entirely different science subjects). These results are 26% and 15% over the majority class baselines, respectively, and 21% and 6% over lexi-

cal entailment baselines based on Glickman et al. (2005).

## 4 Discussion and Future Work

Analyzing the results of reference facet extraction, there are many interesting open linguistic issues in this area. This includes the need for a more sophisticated treatment of adjectives, conjunctions, plurals and quantifiers, all of which are known to be beyond the abilities of state of the art parsers.

Analyzing the dependency parses of 51 of the student answers, about 24% had errors that could easily lead to problems in assessment. Over half of these 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 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 paper and is a significant area of future research. Other key areas of future research involve integrating the representation described here into an ITS and evaluating its impact.

## 5 Conclusion

We presented a novel fine-grained semantic representation and evaluated it in the context of automated tutoring. A significant contribution of this representation is that it will facilitate more precise tutor feedback, targeted to the specific facet of the reference answer and pertaining to the specific level of understanding expressed by the student. This representation could also be useful in areas such as question answering or document summarization, where a series of entailed facets could be composed to form a full answer or summary.

The representation's validity is partially demonstrated in the ability of annotators to reliably annotate inferences at this facet level, achieving substantial agreement (86%, Kappa=0.72) and by promising results in automatic assessment of student answers at this facet level (up to 26% over baseline), particularly given that, in addition to the manual reference answer facet representation, an automatically extracted approximation of the representation was a key factor in the features utilized by the classifier.

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 to the long-term goal of creating ITSs that can truly engage children in natural unrestricted dialog, such as is required to perform high quality student directed Socratic tutoring.

## Acknowledgments

This work was partially funded by Award Number 0551723 from the National Science Foundation.

## References

Briscoe, E., Carroll, J., Graham, J., and Copestake, A. 2002. Relational evaluation schemes. In *Proc. of the Beyond PARSEVAL Workshop at LREC*.

Gildea, D. and Jurafsky, D. 2002. Automatic labeling of semantic roles. *Computational Linguistics*.

Glickman, O, Dagan, I, and Koppel, M. 2005. Web Based Probabilistic Textual Entailment. In *Proc RTE*.

Jordan, P, Makatchev, M, VanLehn, K. 2004. Combining competing language understanding approaches in an intelligent tutoring system. In *Proc ITS*.

Kipper, K, Dang, H, and Palmer, M. 2000. Class-Based Construction of a Verb Lexicon. In *Proc. AAAI*.

Lawrence Hall of Science 2006. Assessing Science Knowledge (ASK), UC Berkeley, NSF-0242510

Leacock, C. 2004. Scoring free-response automatically: A case study of a large-scale Assessment. *Examens*.

Lin, D & Pantel, P. 2001. Discovery of inference rules for Question Answering. In *Natl. Lang. Engineering*.

Nielsen, R, Ward, W, and Martin, JH. 2008a. Learning to Assess Low-level Conceptual Understanding. In *Proc. FLAIRS*.

Nielsen, R, Ward, W, Martin, JH and Palmer, P. 2008b. Annotating Students' Understanding of Science Concepts. In *Proc. LREC*.

Nivre, J, Hall, J, Nilsson, J, Eryigit, G and Marinov, S. 2006. Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines. In *Proc. CoNLL*.

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