A Corpus of Fine-Grained Entailment Relations

Rodney D. Nielsen and Wayne Ward
Center for Spoken Language Research
Institute of Cognitive Science
Department of Computer Science
University of Colorado, Boulder
Rodney.Nielsen, Wayne.Ward@Colorado.edu

Abstract

This paper describes on-going efforts to annotate a corpus of almost 16000 answer pairs with an estimated 69000 fine-grained entailment relationships. We illustrate the need for more detailed classification than currently exists and describe our corpus and annotation scheme. We discuss early statistical analysis showing substantial inter-annotator agreement even at the fine-grained level. The corpus described here, which is the only one providing such detailed annotations, will be made available as a public resource later this year (2007). This is expected to enable application development that is currently not practical.

1 Introduction

Determining whether the propositions in one text fragment are entailed by those in another fragment is important to numerous NLP applications. Consider an intelligent tutoring system (ITS), where it is critical for the tutor to assess which specific facets of the desired or reference answer are entailed by the student's answer. Truly effective interaction and pedagogy is only possible if the automated tutor can assess this entailment at a relatively fine level of detail (c.f. Jordan et al., 2004).

The PASCAL Recognizing Textual Entailment (RTE) challenge (Dagan et al., 2005) has brought the issue of textual entailment before a broad community of researchers in a task independent fashion. This task requires systems to make simple yes-no judgments as to whether a human reading a text *t* of one or more full sentences would typically consider a second, hypothesis, text *h* (usually one full sentence) to most likely be true. This paper discusses some of the extensions necessary to this scheme in order to satisfy the requirements of an ITS and provides a preliminary report on our efforts to produce an annotated corpus applying some of these additions to children's answers to science questions.

We first provide a brief overview of the RTE challenge task and a synopsis of answer assessment technology within existing ITSs and large scale assessment applications. We then detail some of the types of changes required in order to facilitate more effective pedagogy. We provide a report on our work in this direction and describe a corpus we are annotating with fine-grained entailment information. Finally, we discuss future direction and the relevance of this annotation scheme to other applications such as question answering.

2 Prior Work

2.1 RTE Challenge Task

Example 1 shows a typical *t*-*h* pair from the RTE challenge. The task is to determine whether typically a reader would say that *h* is most likely true having read *t*. The system output is a simple yes or no decision about this entailment – in this example, the decision is *no* – and that is similarly the extent to which training data is annotated. There is no indication of whether some facets of, the potentially quite long, *h* are addressed (as they are in this case) in *t* or conversely, which facets are not discussed or are explicitly contradicted.

(1) <t>At an international disaster conference in Kobe, Japan, the

Proceedings of the Workshop on Textual Entailment and Paraphrasing, pages 28–35,
Prague, June 2007.

```
U.N. humanitarian chief said the
United Nations should take the lead
in creating a tsunami early-warning
system in the Indian Ocean.</t>
<h>Nations affected by the Asian
tsunami disaster have agreed the UN
should begin work on an early warning
system in the Indian Ocean.</h>
```

However, in the third RTE challenge, there is an optional pilot task[1] that begins to address some of these issues. Specifically, they have extended the task by including an unknown label, where *h* is neither entailed nor contradicted, and have requested justification for decisions. The form that these justifications will take has been left up to the groups participating, but could conceivably provide some of the information about which specific facets of the hypothesis are entailed, contradicted and unaddressed.

2.2 Existing Answer Assessment Technology

Effective ITSs exist in the laboratory producing learning gains in high-school, college, and adult subjects through text-based dialog interaction (e.g., Graesser et al., 2001; Koedinger et al., 1997; Peters et al., 2004, VanLehn et al., 2005). However, 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 (Aleven et al., 2001; Graesser et al., 2001; Jordan et al., 2004; Peters et al., 2004; Roll et al., 2005; Rosé et al., 2003). However, they are developing very domain-dependent approaches, requiring a significant investment in hand-crafted logic representations, parsers, knowledge-based ontologies, and or dialog control mechanisms. Simply put, these domain-dependent techniques will not scale to the task of developing general purpose ITSs and will never enable the long-term goal of effective unconstrained interaction with learners or the pedagogy that requires it.

There is also a small, but growing, body of research in the area of scoring free-text responses to short answer questions (e.g., Callear et al., 2001; Leacock, 2004; Mitchell et al., 2003; Pullman, 2005; Sukkarieh, 2005). Shaw (2004) and Whittington (1999) provide reviews of some of these approaches. Most of the systems that have been implemented and tested are based on Information Extraction (IE) techniques (Cowie & Lehnert, 1996). They hand-craft a large number of pattern rules, directed at detecting the propositions in common correct and incorrect answers. In general, short-answer free-text response scoring systems are designed for large scale assessment tasks, such as those associated with the tests administered by ETS. Therefore, they are not designed with the goal of accommodating dynamically generated, previously unseen questions. Similarly, these systems do not provide feedback regarding the specific aspects of answers that are correct or incorrect; they merely provide a raw score for each question. As with the related work directed specifically at ITSs, these approaches all require in the range of 100-500 example student answers for each planned test question to assist in the creation of IE patterns or to train a machine learning algorithm used within some component of their solution.

3 The Necessity of Finer-grained Analysis

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

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. I use the word facet throughout this paper to generically refer to some part of a text'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, (i.e., does the student understand the reference answer in its entirety or is there some unspecified part of it that we are unsure whether the student understands), we instead break the reference answer

[1] http://nlp.stanford.edu/RTE3-pilot/

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.[2] The following illustrates how a simple reference answer (2) is decomposed into the answer facets (2a-d) derived from its dependency parse and (2a'-d') provide a gloss of each facet's meaning. As can be seen in 2b and 2c, the dependencies are augmented by thematic roles (Kipper et al., 2000) (e.g., Agent, Theme, Cause, Instrument…) produced by a semantic role labeling system (c.f., Gildea and Jurafsky, 2002). The facets also include those semantic role relations that are not derivable from a typical dependency tree. For example, in the sentence *"As it freezes the water will expand and crack the glass"*, *water* is not a modifier of *crack* in the dependency tree, but it does play the role of Agent in a shallow semantic parse.

(2) A long string produces a low pitch.

(2a) NMod(string, long)
(2b) Agent(produces, string)
(2c) Product(produces, pitch)
(2d) NMod(pitch, low)

(2a′) There is a long string.
(2b′) The string is producing something.
(2c′) A pitch is being produced.
(2d′) The pitch is low.

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. Did the student contradict the facet? Did they express a related concept that indicates a misconception? Did they leave the facet unaddressed? Can you assume that they understand the facet even though they did not express it, since it was part of the information given in the question? It is clear that, in addition to breaking the reference answer into fine-grained facets, it is also necessary to break the annotation into finer levels in order to specify more clearly the relationship between the student's answer and the reference answer aspect.

There are many other issues that the system must know to achieve near optimal tutoring, some of which are mentioned later in the discussion section, but these two – breaking the reference answer into fine-grained facets and utilizing more expressive annotation labels – are the emphasis of this effort.

4 Current Annotation Efforts

This section describes our current efforts in annotating a corpus of answers to science questions from elementary school students.

4.1 Corpus

Lacking data from a real tutoring situation, we acquired data gathered from 3rd-6th grade students in schools utilizing the Full Option Science System (FOSS). Assessment is a major FOSS research focus, of which the Assessing Science Knowledge project is a key component.[3] The FOSS project has developed sixteen science teaching and learning modules targeted at grades 3-6, as shown in Table 1. The ASK project created assessments for each of these modules, including multiple choice, fill in the blank, free response, and somewhat lengthy experimental design questions. We reviewed these questions and selected about 290 free response questions that were in line with the objectives of this research project, specifically we selected questions whose expected responses ranged in length from moderately short verb phrases to a few sentences, that could be assessed objectively, and that were not too open ended. Table 2 shows a

[2] 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).

[3] "FOSS is a research-based science program for grades K–8 developed at the Lawrence Hall of Science, University of California at Berkeley with support from the National Science Foundation and published by Delta Education. FOSS is also an ongoing research project dedicated to improving the learning and teaching of science."
Assessing Science Knowledge (ASK) is "designed to define, field test, and validate effective assessment tools and techniques to be used by grade 3–6 classroom teachers to assess, guide, and confirm student learning in science."
http://www.lawrencehallofscience.org/foss/

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

Table 1 FOSS / ASK Learning and Assessment Modules by Area and Grade

HB	Q: Dancers need to be able to point their feet. The tibialis is the major muscle on the front of the leg and the gastrocnemius is the major muscle on the back of the leg. Describe how the muscles in the front and back of the leg work together to make the dancer's foot point. R: The muscle in the back of the leg (the gastrocnemius) contracts and the muscle in the front of the leg (the tibialis) relaxes to make the foot point. A: The back muscle and the front muscle stretch to help each other pull up the foot.
ST	Q: Why is it important to have more than one shelter in a crayfish habitat with several crayfish? R: Crayfish are territorial and will protect their territory. The shelters give them places to hide from other crayfish. [Crayfish prefer the dark and the shelters provide darkness.] A: So all the crayfish have room to hide and so they do not fight over them.
ME	Q: Lee has an object he wants to test to see if it is an insulator or a conductor. He is going to use the circuit you see in the picture. Explain how he can use the circuit to test the object. R: He should put one of the loose wires on one part of the object and the other loose wire on another part of the object (and see if it completes the circuit). A: You can touch one wire on one end and the other on the other side to see if it will run or not.
PS	Q: Kate said: "An object has to move to produce sound." Do you agree with her? Why or why not? R: Agree. Vibrations are movements and vibrations produce sound. A: I agree with Kate because if you talk in a tube it produce sound in a long tone. And it vibrations and make sound.
WA	Q: Anna spilled half of her cup of water on the kitchen floor. The other half was still in the cup. When she came back hours later, all of the water on the floor had evaporated but most of the water in the cup was still there. (Anna knew that no one had wiped up the water on the floor.) Explain to Anna why the water on the floor had all evaporated but most of the water in the cup had not. R: The water on the floor had a much larger surface area than the water in the cup. A: Well Anna, in science, I learned that when water is in a more open are, then water evaporates faster. So, since tile and floor don't have any boundaries or wall covering the outside, the water on the floor evaporated faster, but since the water in the cup has boundaries, the water in the cup didn't evaporate as fast.
EM	Q: You can tell if a rock contains calcite by putting it into a cold acid (like vinegar). Describe what you would observe if you did the acid test on a rock that contains this substance. R: Many tiny bubbles will rise from the calcite when it comes into contact with cold acid. A: You would observe if it was fizzing because calcite has a strong reaction to vinegar.

Table 2 Sample Qs from FOSS-ASK with their reference (R) and an example student answer (A).

few questions that are representative of those selected for inclusion in the corpus, along with their reference answers and an example student answer for each. Questions without at least one verb phrase were rejected because they were assumed to be more trivial and less interesting from the research perspective. Examples of such questions along with their reference answers and an example student response include: Q: *Besides air, what (if anything) can sound travel through?* Reference Answer: *Sound can also travel through liquids and solids. (Also other gases.)* Student Answer: *A screen door*. Q: *Name a property of the sound of a fire engine's siren*. Reference Answer: *The sound is very loud. OR The sound changes in pitch.* Student Answer: *Annoying*. An example of a free response item that was dropped because it was too open ended is: *Design an investigation to find out a plant's range of tolerance for number of hours of sunlight per day. You can use drawings to help explain your design*.

We generated a corpus from a random sample of the kids' handwritten responses to these questions. The only special transcription instructions were to

fix spelling errors (since these would be irrelevant in a spoken dialog environment), but not grammatical errors (which would still be relevant), and to skip blank answers and non-answers similar in nature to *I don't know* (since these are not particularly interesting from the research perspective).

Three modules were designated as the test set (Environments, Human Body, and Water) and the remaining 13 modules will be used for development and training of classification systems. We judged the three test set modules to be representative of the entire corpus in terms of difficulty and appropriateness for the types of questions that met our research interests. We transcribed the responses of approximately 40 randomly selected students for each question in the training set and 100 randomly selected students for each question in the test set. In order to maximize the diversity of language and knowledge represented by the training and test datasets, random selection of students was performed at the question level rather than using the same students' answers for all of the questions in a given module. However, in total there were only about 200 children that participated in any individual science module assessment, so there is still moderate overlap in the students from one question to another within a given module. On the other hand, each assessment module was given to a different group of kids, so there is no overlap in students between modules. There are almost 60 questions and 5700 student answers in the test set, comprising approximately 20% of all of the questions utilized and 36% of the total number of transcribed student responses. In total, including test and training datasets, there are nearly 16000 student responses.

4.2 Annotation

The answer assessment annotation described in this paper is intended to be a step toward specifying the detailed semantic understanding of a student's answer that is required for an ITS to interact effectively with a learner. With that goal in mind, annotators were asked to consider and annotate according to what they would want to know about the student's answer if they were the tutor (but a tutor that for some reason could not understand the unstructured text of the student's answer). The key exception here is that we are only annotating a student's answer in terms of whether or not it accurately and completely addresses the facets of the reference (desired or correct) answer. So, if the student also discusses concepts not addressed in the reference answer, we will not annotate those points regardless of their quality or accuracy.

Each reference answer in the corpus is decomposed into its constituent facets. Then each student answer is annotated relative to the facets in the corresponding reference answer. As described earlier, the reference answer facets are roughly extracted from the relations in a syntactic dependency parse (c.f., Nivre and Scholz, 2004) and a shallow semantic parse (Gildea and Jurafsky, 2002). These are modified slightly to either eliminate most function words or incorporate them into the relation labels (c.f., Lin and Pantel, 2001). Example 3 illustrates the decomposition of one of the reference answers into its constituent parts along with their glosses.

```
(3) The string is tighter, so the
pitch is higher.

(3a)  Is(string, tighter)
(3a′) The string is tighter.
(3b)  Is(pitch, higher)
(3b′) The pitch is higher.
(3c)  Cause(3b, 3a)
(3c′) 3b is caused by 3a
```

The annotation tool lists the reference answer facets that students are expected to address. Both a formal relational representation and an English-like gloss of the facet are displayed in a table, one row per facet. The annotator's job is to label each of those facets to indicate the extent to which the student addressed it. We settled on the eight annotation labels noted in Table 3. Descriptions of where each annotation label applies and some of the most common annotation issues were detailed with several examples in the guidelines and are only very briefly summarized in the remainder of this subsection.

Example 4 shows a student answer corresponding to the reference answer in example 3, along with its initial annotation in 4a-c and its final annotation in 4a'-c'. It is assumed that the student understands that the pitch is higher (facet 4b), since this is given in the question (*... Write a note to David to tell him why the pitch gets higher rather than lower*) and similarly it is assumed that the student will be explaining what has the causal effect of producing this higher pitch (facet 4c). Therefore, these facets are initialized to `Assumed` by the

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

Expressed: Any facet directly expressed or inferred by simple reasoning
Inferred: Facets inferred by pragmatics or nontrivial logical reasoning
Contra-Expr: Facets directly contradicted by negation, antonymous expressions and their paraphrases
Contra-Infr: Facets contradicted by pragmatics or complex reasoning
Self-Contra: Facets that are both contradicted and implied (self contradictions)
Diff-Arg: The core relation is expressed, but it has a different modifier or argument
Assumed: The system assigns this label, which is changed if any of the above labels apply
Unaddressed: Facets that are not addressed at all by the student's answer

Table 3 Facet Annotation Labels

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

(4a) Is(string, tighter), ---
(4b) Is(pitch, higher), Assumed
(4c) Cause(4b, 4a), Assumed

(4a′) Is(string, tighter), Diff-Arg
(4b′) Is(pitch, higher), Expressed
(4c′) Cause(4b, 4a), Expressed
```

The `Self-Contra` annotation is used in cases like the response in example 5, where the student simultaneously expresses the contradictory notions that the string is tighter and that there is less tension.

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

(5a) Is(string, tighter), Self-Contra
```

There is no compelling reason from the perspective of the automated tutoring system to differentiate between `Expressed` and `Inferred` facets, since in either case the tutor can assume that the student understands the concepts involved. However, from the systems development perspective there are three primary reasons for differentiating between these facets and similarly between facets that are contradicted by inference versus more explicit expression. The first reason is that most statistical machine learning systems today cannot hope to detect very many pragmatic inferences and including these in the training data is likely to confuse the algorithm resulting in worse performance. Having separate labels allows one to remove the more difficult inferences from the training data, thus eliminating this problem. The second rationale is that systems hoping to handle both types of inference might more easily learn to discriminate between these opposing classifications if the classes are distinguished (for algorithms where this is not the case, the classes can easily be combined automatically). Similarly, this allows the possibility of training separate classifiers to handle the different forms of inference. The third reason for separate labels is that it facilitates system evaluation, including the comparison of various techniques and the effect of individual features.

Example 6 illustrates an example of a student answer with the label `Inferred`. In this case, the decision requires pragmatic inferences, applying the Gricean maxims of Relation, be relevant – why

would the student mention vibrations if they did not know they were a form of movement – and Quantity, do not make your contribution more informative than is required (Grice, 1975).

```
(6) Q: Kate said: "An object has to
move to produce sound." Do you agree
with her? Why or why not?
Ref Ans: "Agree. Vibrations are move-
ments and vibrations produce sound."
Student Answer: Yes because it has to
vibrate to make sounds.

(6b) Is(vibration, movement), Inferred
```

Annotators are primarily students of Education and Linguistics and require moderate training on the annotation task. The annotated reference answers are stored in a stand-off markup in xml files, including an annotated element for each reference answer facet.

4.3 Inter-Annotator Agreement Results

The results reported here are preliminary, based on the first two annotators, and must be viewed under the light that we have not yet completed annotator training. We report results under three label groupings: (1) `All-Labels`, where all labels are left separate, (2) `Tutor-Labels`, where `Expressed`, `Inferred` and `Assumed` are combined as are `Contra-Expr` and `Contra-Infr`, and (3) `Yes-No`, which is a two-way division, `Expressed`, `Inferred` and `Assumed` versus all other labels.

Agreement on `Tutor-Labels` indicates the benefit to the tutor, since it is relatively unimportant to differentiate between the types of inference required in determining that the student understands a reference answer facet (or has contradicted it). We evaluated ***mid-training*** inter-annotator agreement on a random selection of 15 answers from each of 14 Physics of Sound questions, totaling 210 answers and 915 total facet annotations. Mid-training agreement on the `Tutor-Labels` is 87.4%, with a Kappa statistic of 0.717 corresponding with substantial agreement (Cohen, 1960). Inter-annotator agreement at mid-training is 81.1% on `All-Labels` and 90.1% on the binary `Yes-No` decision. These also have Kappa statistics in the range of substantial agreement.

The distribution of the 915 annotations is shown in Table 4. It is somewhat surprising that this science module had so few contradictions, just 2.7% of all annotations, particularly given that many of the questions seem more likely to draw contradictions than unaddressed facets (e.g., many ask about the effect on pitch and volume, typically eliciting one of two possible responses). An analysis of the inter-annotator confusion matrix indicates that the most probable disagreement is between `Inferred` and `Unaddressed`. The second most likely disagreement is between `Assumed` and `Expressed`. In discussing disagreements, the annotators almost always agree quickly, reinforcing our belief that we will increase agreement significantly with additional training.

Label	Count	%	Count	%
`Expressed`	348	38.0	657	71.8
`Inferred`	51	5.6		
`Assumed`	258	28.2		
`Contra-Expr`	21	2.3	25	2.7
`Contra-Infr`	4	0.4		
`Self-Contra`	1	0.1	1	0.1
`Diff-Arg`	33	3.6	33	3.6
`Unaddressed`	199	21.7	199	21.7

Table 4 Distribution of classifications (915 facets)

5 Discussion and Future Work

The goal of our fine-grained classification is to enable more effective tutoring dialog management. The additional labels facilitate understanding the type of mismatch between the reference answer and the student's answer. Breaking the reference answer down into low-level facets enables the tutor to provide feedback relevant specifically to the appropriate facet of the reference answer. In the question answering domain, this facet-based classification would allow systems to accumulate entailing evidence from a variety of corroborating sources and incorporate answer details that might not be found in any single sentence. In other applications outside of the tutoring domain, this fine-grained classification can also facilitate more directed user feedback. For example, both the additional classifications and the break down of facets can be used to justify system decisions, which is the stated goal of the pilot task at the third RTE challenge.

The corpus described in this paper, which will be released later this year (2007), represents a substantial contribution to the entailment community, including an estimated 69000 facet entailment annotations. By contrast, three years of RTE challenge data comprise fewer than 4600 entailment

annotations. More importantly, this is the only corpus that provides entailment information at the fine-grained level described in this paper. This will enable application development that was not practical previously.

Future work includes training machine learning algorithms to perform the classifications described in this paper. We also plan to annotate other aspects of the students' understanding that are not direct inferences of reference answer knowledge. Consider example (4), in addition to the issues already annotated, the student contradicts a law of physics that they have surely encountered elsewhere in the text, specifically that longer strings produce lower, not higher, pitches. Under the current annotation scheme this is not annotated, since it does not pertain directly to the reference answer which has to do with the effect of string tension. In other annotation plans, it would be very useful for training learning algorithms if we provide an indication of which student answer facets played a role in making the inferences classified.

Initial inter-annotator agreement results look promising, obtaining substantial agreement according to the Kappa statistic. We will continue to refine our annotation guidelines and provide further training in order to push the agreement higher on all classifications.

Acknowledgement

We would like to thank Martha Palmer for valuable advice on this annotation effort.

References

Aleven V, Popescu O, & Koedinger K. (2001) A tutorial dialogue system with knowledge-based understanding and classification of student explanations. *IJCAI WS knowledge & reasoning in practical dialogue systems*

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

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational & Psych Measurement*. 20:37-46.

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

Dagan, Ido, Glickman, Oren, and Magnini, Bernardo. (2005). The PASCAL Recognizing Textual Entailment Challenge. In *1st RTE Challenge Workshop*.

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

Graesser, A.C., Hu, X., Susarla, S., Harter, D., Person, N.K., Louwerse, M., and Olde, B. (2001). AutoTutor: An Intelligent Tutor and Conversational Tutoring Scaffold. In *10th ICAI in Education*, 47-49.

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

Jordan, P. W., Makatchev, M., & VanLehn, K. (2004). Combining competing language understanding approaches in an intelligent tutoring system. In *7th ITS*.

Kipper, K, Dang, H, & Palmer, M. (2000). Class-Based Construction of a Verb Lexicon. *AAAI 17th NCAI*

Koedinger, K.R., Anderson, J.R., Hadley, W.H. & Mark, M.A. (1997). Intelligent tutoring goes to school in the big city. *Intl Jrnl of AI in Ed, 8*, 30-43.

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

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

Mitchell, T. Aldridge, N., and Broomhead, P. (2003). Computerized marking of short-answer free-text responses. In *29th IAEA*.

Nivre, J. and Scholz, M. (2004). Deterministic Dependency Parsing of English Text. In *Proc COLING*.

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

Pulman S.G. & Sukkarieh J.Z. (2005). Automatic Short Answer Marking. *ACL WS Bldg Ed Apps using NLP*.

Roll, I, Baker, R, Aleven, V, McLaren, B, & Koedinger, K. (2005). Modeling Students' Metacognitive Errors in Two Intelligent Tutoring Systems. In *UM* 379–388

Rosé, P. Roque, A., Bhembe, D. & VanLehn, K. (2003). A hybrid text classification approach for analysis of student essays. In *Bldg Ed Apps using NLP*

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

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

VanLehn, K., Lynch, C., Schulze, K. Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2005). The Andes physics tutoring system: Five years of evaluations. In *12th ICAI in Ed*

Whittington, D., Hunt, H. (1999). Approaches to the Computerised Assessment of Free-Text Responses. *Third ICAA*.