TAC 2008 CLEAR RTE System Report: Facet-based Entailment

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

Abstract

This paper describes the CLEAR team's submission to the 2008 Text Analysis Conference under the Recognizing Textual Entailment track. The system breaks text fragments down into fine-grained semantic facets and performs entailment recognition on these. We show that, in the relevant subset of the data, we can achieve 90% accuracy in pinpointing the specific facet of a hypothesis that is not entailed. We also provide an error analysis based on the facets of hypotheses most likely to have led to their misclassification.

1 Introduction

Recognizing textual entailment (RTE) can be beneficial in a wide variety of applications such as Question Answering (QA), Document Summarization and Intelligent Tutoring Systems (ITSs). Many of these applications could benefit from a more fine-grained analysis of the entailment relations between the two text fragments. For example, in Multiple Document Summarization, if two documents imply the same fine-grained semantic facet, it can be added to the summary with greater confidence and, conversely, if a second document contradicts the first on a facet, that facet can be omitted or tagged for an analyst to review. Similarly, in an ITS, it is important for the system to detect *specifically* where and in what way a student's answer varies from the desired reference answer.

Rather than have a single entailed versus not-entailed assessment of the hypothesis text, *h*, as a whole, we instead break *h* 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 *h*. Breaking *h* down into fine-grained facets permits a more focused assessment of a student's response in an ITS or a more detailed analysis of a potential answer to a question in a QA system.

In this paper, we describe our approach to automatically decomposing text into fine-grained facets, describe our system to automatically classify facets as entailed or not entailed by the reference text, *t*, and provide an error analysis of the system at the facet level. This is done in the context of the RTE track of TAC 2008. In this track, systems are given a reference text, *t*, and corresponding hypothesis text, *h*, and the objective is to classify *h* as being fully Entailed by *t*, Contradicted by *t*, or as having Unknown veracity. Examples of such RTE *t*-*h* pairs, where the truth of *h* cannot be determined from *t*, follow.

(18.*t*) The victims' families, as well as women who survived Michel Fourniret's alleged attacks, sat opposite the accused and his wife Monique Olivier on the first day of the trial for the kidnap, rape and murder of seven young women and girls.
(18.*h*) Michel Fourniret was sentenced to life imprisonment.

(179.*t*) Those expecting higher beef and pork prices aren't ready to say by how much. Mark Schultz, chief analyst at Northstar Commodity Investment in Minneapolis, predicted that beef prices will rise substantially. (For the latest commodity prices, go here.)
(179.*h*) Corn prices increase.

In example 18, we would like to specifically determine that there was no *imprisonment sentence* indicated in *t* and in example 179, we want to indicate that it is not *corn prices* that have increased.

2 Facet-based Representation

We automatically decomposed each hypothesis, *h*, into fine-grained facets, roughly extracted from the relations in a syntactic dependency parse. However, we use the word *facet* to refer to any fine-grained component of the semantics. These facets are the basis for assessing whether *h* as a whole is entailed and, if not, what specific area lacks entailment. See (Nielsen et al., 2008b) for details on extracting the facets; here we simply sketch the transformation into the final representation.

Figure 1 shows the original dependency parse for 18.*h* along with the final automatically extracted facet-based representation. In this example, relatively few changes were made to the original parse. In general, the system reattaches auxiliary verbs and their modifiers to the associated main verbs. It incorporates prepositions and copulas into the dependency relation labels, and similarly appends negation terms onto the associated dependency relations. These modifications increase the likelihood that terms carrying significant semantic content are joined by dependencies that are utilized in feature extraction. For example, the dependencies vmod[1](*sentenced*, *to*) and pmod[2](*to*, *imprisonment*) provide little semantic value over the single content word in each dependency. Whereas, the facet vmod_to(*sentenced*, *imprisonment*) carries more semantic value.

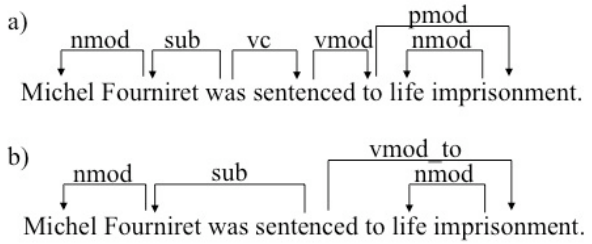

Figure 1. RTE ex. 18 hypothesis a) dependency parse and b) facet representation

Example 1 below, taken from our corpus of grade 3-6 student responses to science questions (Nielsen et al., 2008a) presents a more involved transformation and illustrates the facets derived from its dependency parse (shown in Figure 2), along with their glosses. These facets represent the fine-grained knowledge the student is expected to address in their response to the associated Assessing Science Knowledge (Lawrence Hall of Science, 2006) assessment question.

(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.

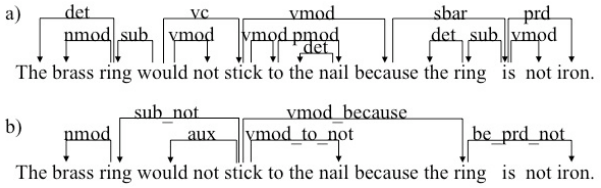

Figure 2. Hypothesis a) dependency parse and b) facet representation

Unlike with the RTE hypotheses, where the facet representation was automatically generated, in the children's corpus we manually extracted the facets from each question's reference answer. Typical facets, as in (1a), are derived directly from a dependency parse, in this case retaining its dependency type label, NMod. 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.

[1] Verb Modifier
[2] Preposition Modifier

3 The Entailment System

A high level description of the entailment procedure is as follows. The system first decomposes *h* into its constituent facets as described in section 2. Then for each facet in *h*, we extract features indicative of whether *t* entails that facet. We trained a machine learning classifier on our corpus of children's answers to science questions, which is labeled to indicate whether each facet of a desired reference answer is entailed by the student's answer. We use this classifier to compute entailment probabilities for each facet of *h*. Considering all of the facets in *h*, we then constructed a feature vector from combinations of these probability estimates and features of the corresponding facets. Finally, we trained a classifier on past years' RTE datasets and used it to classify the RTE4 test examples.

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 and were indexed and searched using Lucene, a publicly available Information Retrieval tool.

Before extracting features, we automatically generate dependency parses of *h* and *t* using MaltParser (Nivre et al., 2006). These parses are then automatically modified as sketched in section 2. We reattach auxiliary verbs and their modifiers to the associated main verbs. We incorporate prepositions and copulas into the dependency relation labels, and similarly append negation terms onto the associated dependency relations. These modifications increase the likelihood that terms carrying significant semantic content are joined by dependencies that are utilized in feature extraction.

3.2 Facet Entailment

We investigated a variety of linguistic features and chose to utilize the features summarized in Table 1, 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.

Lexical Features
Gov/Mod_MLE: The lexical entailment probabilities (LEPs) for the facet governor and modifier following (Glickman et al., 2005; c.f., Turney, 2001). The LEP of a hypothesis word *w* is defined as: (1) $LEP(w) = \max_{v \in t}(n_{w,v}/n_v)$, where *v* is a word in *t*, n_v is the # of docs (see section 3.1) containing *v*, and $n_{w,v}$ is the # of docs where *w* & *v* cooccur.
Gov/Mod_Match: True if the Gov (Mod) stem has an exact match in *t*.
Subordinate_MLEs: The lexical entailment probabilities for the primary constituent facets' Govs and Mods when the facet represents a relation between higher-level propositions.
Syntactic Features
Gov/Mod_POS: POS tags for the facet's Gov and (Mod).
Facet/AlignedDep_Reltn: The labels of the facet and aligned dependency in *t* – alignments were based on co-occurrence MLEs as with words, (i.e., they estimate the likelihood of seeing the *h* dependency in a document given it contains the *t* dependency – replace words with dependencies in equation 1 above).
Dep_Path_Edit_Dist: The edit distance between the dependency path connecting the facet's Gov and Mod (always a single step for the automatically generated RTE facets, but not necessarily a single step in our children's corpus due to parser errors, etc.) and the path connecting the aligned terms in *t*. 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).
Other Features
Consistent_Negation: True if the *h* facet and aligned *t* dependency path had the same number of negations.
RA_CW_cnt: The number of content words (non-function words) in *h*.

Table 1. Machine Learning Features

Our research is in the domain of Intelligent Tutoring Systems, where we have annotated a corpus

of grade 3-6 student answers to science questions. This corpus consists of 287 questions, approximately 15,400 total student answers, and nearly 146K fine-grained facet entailment annotations. We used a subset of this data (primarily those facets that were expressed or paraphrased by the student answer and those left unaddressed by the student answer) to train our facet-level entailment classifier.

We evaluated several machine learning algorithms (rules, trees, boosting, ensembles and an SVM) and C4.5 (Quinlan, 1993) marginally achieved the best results in cross validation on the training data. A thorough analysis of the impact of the classifier chosen has not been completed at this time. We then used this classifier, trained on the children's data, to classify each facet in *h* as being entailed or not entailed by *t*. The probability estimates from these classifications were then combined as described in the next section to generate feature vectors for predicting the overall hypothesis entailment.

3.3 Hypothesis Entailment

We participated only in the two-way (Entailment versus No Entailment) classification task, since our current features (at the facet level and the overall *t-h* pair level) are not yet designed to detect contradictions.

Our features for the entailment classification of *h*, consist of the average, geometric mean and worst entailment probabilities calculated for the individual facets in *h*; similar calculations for the facets' governors, the modifiers, and the path edit distances; the proportion of governors and modifiers that had an exact stem match in *t*; the proportion of governors and modifiers that had non-zero co-occurrence statistics; the proportion of facets where both terms had exact matches; the proportion where either had an exact match; the proportion of aligned paths where negations were consistent with the hypothesis facet; the part-of-speech tags for the governor, modifier, and their aligned terms for the facet with the worst entailment probability; and the number of content words and facets in *h*.

We created a training set for the final classifier from all of the data in the prior RTE challenges. Each of the three runs that we submitted was determined by combining the output of a wide variety of learning algorithms, including rules, trees, boosting, ensembles and an SVM. The classification of *h* in the first two runs was based on a majority vote of the classifiers and by averaging the entailment probability estimates of the classifiers. The third run, which marginally provided the best results, used a Stacking classifier to combine the results of the constituent classifiers. The results of this third run are presented and analyzed in the remaining sections of the paper.

3.4 System Results

The results for our system are shown in . The columns list accuracy by entailment pair type, and the rows represent results broken down by task.

Task	All Examples	Entailed Pairs	Non-Entailed Pairs
QA	50.0	80.0	20.0
SUM	64.5	66.0	63.0
IR	70.3	77.3	63.3
IE	55.3	86.7	24.0
ALL	60.6	78.4	42.8

Table 2. Classifier Accuracy

4 Discussion and Error Analysis

4.1 Results Discussion

The accuracy across all text-hypothesis pairs was 60.6%. By task, our system performed best overall on the IR task. However there was a dramatic drop in performance for non-entailed *t-h* pairs in the IE and QA tasks. Our system exhibited a similar gap for non-entailed IE and SUM pairs on RTE3 data. If we exclude the bottom two groups our accuracy for RTE4 is 73.3% and for RTE3 81.5%. We believe this disparity reflects less on the task and more on the type of inference needed to correctly classify an example. A large proportion of the examples in our bottom performing groups required deep logical inference. For example:

(946.*t*) Beijing has threatened a military attack if the Taiwan independence is declared. The two sides split amid civil war in 1949 when the Communists established the People;s Republic and the Nationalist Party, or Kuomintang, moved the original government to Taiwan where they maintained the Republic of China, which Beijing regards as

defunct. President Chen maintains there are two countries.

(946.*h*) Taiwan has been independent since 1949.

To recognize that the *h* above is not entailed requires understanding from the *t* that *Taiwan's independence* is disputed and thus cannot be definitively labeled as entailed and thus should be labeled unknown. Perhaps the organizers of future RTE tracks can consider labeling data with inference type in addition to task type.

Because our system decomposes hypotheses into fine-grained semantic facets, it is well suited to identifying what parts of a hypothesis are more likely to cause it to not be entailed. This lower-level breakdown is useful in ITS applications to provide justification for why a student's response is incorrect. For this reason, we concentrate on analyzing how well the system can identify what causes an *h* to be predicted as *not entailed*. We hypothesize that the facet-based approach will provide an effective method for this identification.

From the RTE4 test set, we took a random sample of 100 *t-h* pairs labeled *unknown*, which were correctly identified as *not entailed* by our system. Pairs labeled *contradiction* are excluded from this analysis since our system and its features were not tuned to recognize contradictions. For each member of the sample, we recorded the lowest probability of facet entailment over all facets in the *h*. We then looked at the facet with the lowest entailment probability for each pair in the sample set and hand annotated whether or not it individually was entailed by the text. Ninety-two of the selected pairs had a facet that was not entailed. Of these pairs, only 10% were incorrectly labeled entailed. In other words, our system can provide justification for why an *h* is not entailed with 90% accuracy.

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. Several randomly selected text-hypothesis pairs were analyzed to look for patterns in the types of errors the system makes. For the analysis, pairs labeled *unknown* or *contradiction* were both treated as *not entailed*.

We discuss Entailed pairs in the next section of the paper and Not Entailed pairs in the subsequent section.

4.3 Errors in Entailed Pairs

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 facets with the lowest probability of entailment were entailed and what relationships exist between the text and the hypothesis.

We analyzed 100 random examples of errors where annotators labeled the hypothesis Entailed and the system labeled it Not Entailed. Out of these 100 examples, eight looked as if they were incorrectly annotated and three appeared to be due to a bug in the system. We group the potential error factors seen in the data, listed in order of frequency, according to issues associated with paraphrases, pragmatics and logical inference, and preprocessing. In the following paragraphs, these groups are broken down for a more detailed analysis.

Paraphrase issues, taken broadly, are subdivided into four main categories: Phrase-based paraphrases, lexical substitution, syntactic alternation and coreference. Our results in this area are in line with the analysis of Bar-Haim et al. (2005). The largest category of paraphrase error involves phrase-based paraphrases 26 errors. Examples of this category include: *not telling the truth* for *lying*, *not fully available to the public* for *private*, and *outside the solar system* for *extrasolar*.

The next largest category of paraphrase error is simple lexical substitution (consisting of synonymy, hypernymy, hyponymy, meronymy, derivational changes, and other lexical paraphrases). Roughly half of these relationships should be detectable using broad coverage lexical resources – for example, substituting *attack* for *assault*, *happened* for *occurred*, *faking* for *pretending*, and *raised* for *reared*. However, many of these lexical paraphrases are not necessarily associated in lexical resources such as WordNet. For example, in the substitution of *training* for *programs* these terms are only connected at the top of the WordNet hierarchy at the Synset(psychological feature).

Our analysis suggested that only about X% of errors could be resolved strictly by syntactic analysis. However, see Vanderwende et al. (2005) for

an analysis that suggested as much as 34% of the full RTE1 test set could be handled by recognizing simple syntactic variations.

Whereas coreference errors accounted for nearly 30% of the paraphrase errors when using a similar system for ITS applications (Nielsen 2008b), only 2 errors in this sample could be primarily attributed to lack of coreference resolution. While numerous coreferences are encountered in children's tutoring data, the RTE4 data is more explicit when referring to entities.

Combined, deep logical reasoning and pragmatics were involved in 20% of the issues. Twelve errors came from Pragamatics divided nearly evenly between errors of implicature and errors of presupposition. Examples of implicature include recognizing that saying *Our sister planet* is identical to *Earth's sister planet* unless the phrase is preceded with additional information like *The Martian said*. Presupposition errors imply something is true independent of whether or not the statement is negated. For example, *is captain of the sunken Princess of the Stars* and *is not captain of the Princess of the Stars* both imply *Princess of the Stars is a ship*. Logical inference errors involve higher-level processes or computation, (e.g., to understand that *80 percent* also implies *at least 70 percent* or that *a record jump in oil prices* coupled with *Efficiency, shared technology and the promotion of alternative power sources will be high on the agenda* indicates the existence of an *oil crisis*.

Lastly, preprocessing errors also accounted for 12 errors. Simple data normalization (e.g. *35-minute* to *35 minute* or *delivery man* to *delivery-man*) should resolve the majority of these errors. It is interesting to note, that a third of preprocessing errors were related to named entities where the system was unable to resolve *United Kingdom* to *UK* or *Group of Eight* to *G8*. Again, a simple gloss should alleviate the majority of these issues.

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). In none of the sampled data was the primary cause of failure attributed to a bad parse. 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.

4.4 Errors in Non-Entailed Pairs

One of the biggest sources of errors in non-entailed examples results from ignoring the context of words. Consider the following:

(471.*t*) Barely six months after her marriage with the French President, supermodel Carla Bruni has admitted having problems with her "conservative" hubby Nicolas Sarkozy's "right-wing politics".

(471.*h*) Carli Bruni is the French President.

To make the correct decision the system needs to take into account that Carli Bruni is married to the prime minister and is not actually the prime minister herself.

Many of the errors in Non-Entailed pairs appear to be the result of facets having antonyms which have very similar statistical co-occurrence patterns. Examples of these types of errors include confusing *climb* with *decrease* distance and *convicted* with *acquitted*. However, both of these cases were actually labeled *contradiction* which means a system trained to identify contradictions could potentially avoid this mistake. Similarity in co-occurrence patterns was not limited to antonym pairs. High co-occurence numbers for words like *terrorist* and *attacked* or *won* and *seats* also contributed to incorrect predictions of entailment.

The most common source of error is simply classifying a number of facets as Understood if there is partial lexical similarity and perhaps syntactic similarity as is the case with this example:

(780.*t*) The White House eliminated funding for a service mission the Hubble Space Telescope from its 2006 budget request and directed NASA to focus on deorbiting the spacecraft at the end of its life, according to government and industry sources.

(780.*h*) NASA destroyed the Hubble Space Telescope.

The processes and the more informative features that would be required to handle the errors on the entailed pairs described in the preceding section should allow the learning algorithm to focus on less noisy features and also 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.

5 Conclusion

We described a novel technique to identify the fine-grained semantic units that result in decisions of No Entailment. In the relevant subset of data analyzed, these decisions were made with 90% accuracy. We are currently developing an Intelligent Tutoring System that will incorporate these strategies with the goal of increasing student learning gains by focusing questions and feedback on the specific facets of the reference answer that the student did not adequately address.

Acknowledgements

This work was partially funded by Award Numbers R305B070434 from IES and 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.