Assessing Student Self-Explanations in an Intelligent Tutoring System

Vasile Rus (vrus@memphis.edu)
Department of Computer Science
Institute for Intelligent Systems
The University of Memphis
Memphis. TN 38152

Philip M. McCarthy (pmmccrth@memphis.edu)
Department of Psychology
Institute for Intelligent Systems
The University of Memphis
Memphis. TN 38152

Arthur C. Graesser (a-graesser@memphis.edu)
Department of Psychology
Institute for Intelligent Systems
The University of Memphis
Memphis. TN 38152

Mihai C. Lintean (M.Lintean@memphis.edu)
Department of Computer Science
Institute for Intelligent Systems
The University of Memphis
Memphis. TN 38152

Danielle S. McNamara (d.mcnamara@mail.psyc.memphis.edu)
Department of Psychology
Institute for Intelligent Systems
The University of Memphis
Memphis. TN 38152

Abstract

Research indicates that guided feedback facilitates learning, whether in the classroom or with Intelligent Tutoring Systems (ITS). Improving the accuracy of the evaluation of user input is therefore necessary for providing optimal feedback. This study investigated an automated assessment of students' input that involved a lexico-syntactic (entailment) approach to textual analysis along with a variety of other textual assessment measures. The corpus consisted of 357 student responses taken from a recent experiment with iSTART, an ITS that provides students with self-explanation and reading strategy training. The results of our study indicated that the entailment approach provided the highest single measure of accuracy for assessing input when compared to the other measures in the study. A set of indices working in conjunction with the entailment approach provided the best overall assessments.

Keywords: entailment; intelligent tutoring systems; iSTART; paraphrase; latent semantic analysis.

Introduction

A major challenge for Intelligent Tutoring Systems (ITSs) that incorporate natural language interaction is to accurately evaluate users' contributions and to produce accurate feedback. Available research in the learning sciences indicates that guided feedback and explanation is more effective than simply providing an indication of *rightness* or *wrongness* of student input (Aleven & Koedinger, 2002; Anderson et al., 1989; Kluger & DeLisi, 1996; McKendree, 1990; Sims-Knight & Upchurch, 2001). And the benefits of feedback specifically in ITS are equally evident (Azevedo & Bernard, 1995). This study addresses the challenge of evaluating users' textual input in ITS environments. More specifically, we assess *entailment* evaluations that are generated from a lexico-syntactic computational tool called The Entailer (Rus et al., 2005; Rus, McCarthy, & Graesser, 2006).

Our corpus of natural language input generated from an ITS is student contributions from users of iSTART (Interactive Strategy Training for Active Reading and Thinking; McNamara, Levinstein, & Boonthum 2004), a web based tutoring system that provides students with self-explanation and reading strategy training. The iSTART student statements were sampled from the final phase of iSTART training. During this stage, a pedagogical agent reads sentences from a textbook aloud and asks the student to type a self-explanation of each sentence. The focus of this study is to distinguish two very similar student self-explanation categories: *Topic identification sentences* and *Paraphrases*. This distinction is challenging because the lexicon used for both topic identification and paraphrase tends to largely overlap with the iSTART target sentences. Thus, for the iSTART agent to provide the most appropriate feedback to the student, accurate algorithms are required to successfully interpret the student's input and make this distinction. This study tests various measures for evaluating student input and formulates an algorithm from a combination of successful indices. The algorithm accurately assesses the student input, distinguishing topic sentence type self explanations from paraphrase-type self explanation. Thus, once implemented, iSTART agents will be able to provide more informative feedback to students.

Interactive Strategy Training for Active Reading and Thinking (iSTART)

iSTART provides young adolescent to college-aged students with tutored self-explanation and reading strategy training via pedagogical agents (McNamara et al., 2004). iSTART is designed to improve students ability to self-explain by teaching them to use reading strategies such as

comprehension monitoring, *bridging*, and *paraphrasing*. Following *introduction* and *practice* phases of the iSTART training, the final practice phase has students use reading strategies by typing self-explanations of sentences from science texts. For example, the following sentence, called Text (T), is from a science textbook and the student input, called self-explanation (SE), is reproduced from a recent iSTART experiment. The SE samples in this study are all reproduced as typed by the student.

T: *The largest and most visible organelle in a eukaryotic cell is the nucleus.*
SE: *the nucleusis the center of the cell it contains the ribsome and more.*

Computational Approaches to Text Assessment

Providing appropriate feedback to students concerning self-explanations requires an accurate evaluation of both the meaning and quality of the self-explanation. In order to assess the best measures available, we assessed seven approaches to self explanation evaluation. The algorithms differ in terms of whether they are *word*-based, incorporate syntactic information, or use a combination of both word and syntactic information.

(1) Latent Semantic Analysis (LSA) The ability of LSA (Landauer et al., 2007) to evaluate similarities between texts is based on particular statistical analyses of word-by-text and word-by-word co-occurrence matrices. Essentially, LSA bases semantics on the premise that a word's meaning is related to the kind of words with which it tends to co-occur. Thus, *chair* is closer in meaning to *table*, *sit*, and *easy* than *chair* is to *horse*, because *chair-table*, *chair-sit*, and *easy-chair* co-occurs in texts more often than does *chair-horse.* LSA has an excellent record of success in text comparison analyses (Landauer et al., 2007), but three major problems with LSA reduce its ability to accurately assess short text of the sort commonly encountered in ITS dialogue. First, LSA does not encode word order (syntax). Second, LSA ignores negation. And third, longer sentence pairs tend to be judged by the LSA as more similar because longer texts increase the likelihood of word similarity between word pairs (McCarthy et al., 2007).

(2) The Entailment Index *The Entailment Index*, generated from The Entailer (Rus et al., 2005; Rus et al., 2006), is applied in this study based on previous success in assessing similarity between short dialogue exchanges in natural language environments (McCarthy et al., 2007). The Entailment Index is relatively impervious to the three major challenges of LSA. Because The Entailment Index is a relatively new metric, we describe its calculation in some detail (see Figure 1 for The Entailer's process flow).

Measuring entailment requires assessing whether the meaning of one text, referred to as the *Hypothesis*, or simply H, can be logically inferred from another text, referred to as the *Text*, or simply T (Dagan, Glickman, & Magnini, 2005). In the iSTART context, the Text (T) corresponds to the textbook sentence and the Hypothesis (H) to the Self-Explanation. Our approach to measuring entailment begins with mapping the T (textbook sentence) and H (student self explanation) into a graph representation (Rus et al., 2005). Words are mapped onto vertices (V) and syntactic relations among words are mapped onto edges (E) in the graph.

The mapping process has three phases: *preprocessing*, *dependency graph generation*, and *final graph generation*. In the preprocessing phase, we (a) strip the punctuation from words (tokenization), (b) map morphological variations of words to their base or root form (lemmatization), (c) assign part-of-speech labels to each word (tagging), and (d) identify the inter-relationship of major phrases within the texts (parsing). The second phase is the actual mapping from text to the graph representation. This mapping is based on information from parse trees generated during the parsing process. A parse tree groups words in a sentence into phrases and organizes phrases into hierarchical tree structures from which we can detect syntactic dependencies among concepts. We use Charniak's (Charniak 2000) parser to obtain such parse trees and head detection rules (Magerman, 1994) to obtain the head of each phrase. A dependency tree is generated by linking the head of each phrase to its modifiers. In the third phase, the dependency tree is transformed into a dependency graph by generating remote dependencies such as the dependency between *speak* and *person* in the sentence *I saw the person I spoke to*.

Once graph representations have been obtained, a *graph matching operation* is initialized. This operation evaluates the degree of similarity between graphs. Several variations of graph matching exist, but the *subsumption* model best fits our task. Graph subsumption consists of finding a mapping from the vertices (V) in SE to the vertices in T such that edges (E) among the same two vertices in SE hold among mapped vertices in T. The subsumption algorithm for textual entailment has three major steps: (1) find an isomorphism between the set of vertices of the Hypothesis graph (V_H) and the Text graph (V_T); (2) check whether the labeled edges in H, E_H, have correspondences in E_T; and (3) compute the subsumption score. Step 1 uses a word-matching method and a thesaurus (Miller, 1995) to find all possible synonyms for words in T (Rus et al., 2005). Words in H have different priorities: head words are most important followed by modifiers. Step 2 takes each relation in H and checks its presence in T. In Step 2, we also use relation equivalences among appositions, possessives and linking verbs. Lastly, a normalized score for vertex and edge mapping is computed. The score for the entire entailment is the weighted sum of each individual vertex and edge matching score. The evaluation is structured so as to generate a value that ranges from 0 to 1, with 1 meaning TRUE entailment and 0 meaning FALSE entailment. However, one final stage of the evaluation is then implemented to account for negation: If only one of the text fragments (i.e., H or T) is negated, the entailment decision is

reversed; however, if an even number of negations occur (e.g., both T and H are negated) the decision is retained (double-negation). For example, the Text *Yahoo bought Overture* does *not* entail the Hypothesis *Yahoo did not buy Overture* because even though the Text subsumes the Hypothesis, the presence of negation reverses that decision. (For an extensive review of the components of the Entailer and the evaluation formula, see Rus et al., 2005).

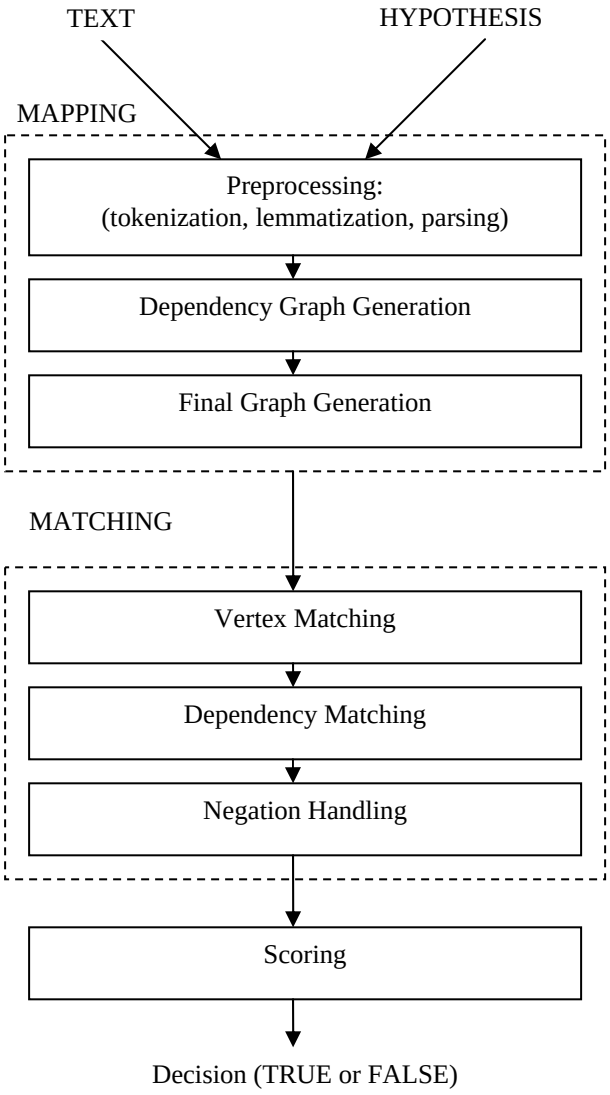

Figure 1: Processing flow in our approach.

(3) Word-Overlap Approach-1 The first word based approach in this study incorporates a simple lexical overlap method: tokenize, lemmatize (using *wnstemm* algorithm in WordNet library), and compute the degree of lexical overlap between the Text and SE. We normalize results by dividing the lexical overlap by the total number of words in the SE. The normalization factor makes the difference in this approach. The values indicate how much the SE is subsumed by the T and not the other way around (entailment approach). This approach is asymmetrical; the values are different if we switch T and SE.

(4) Word-Overlap Approach-2 This second word overlap approach differs from the first in that a cosine value is derived from vectors formed from word co-occurrences. This approach is symmetrical, indicating the degree of similarity between two sentences.

(5) Lemma-Overlap Approach The lemma overlap approach is calculated in the same way as the second word overlap approach, except that lemma co-occurrence rather than simple word co-occurrence is evaluated. For example, the lemma index evaluates *table*/*tables* and *run*/*ran* as the same, whereas the word indices view such pairs as different. Like the second overlap approach, this index is symmetrical, providing a similarity measure between the two sentences.

(6) Synonymy This metric simply adds synonymy to the first word overlap method. The synonymy and word-based approaches are equivalent to The Entailer's lexico-syntactic approach for cases when only the lexical component is used and the syntactic component is ignored. This is an entailment approach.

(7) Syntactic The syntactic approach is equivalent to The Entailer's lexico-syntactic approach when only the syntactic component is used and the lexical component is ignored. The evaluation of the latter allows us to understand the degree to which syntax alone can contribute to entailment. For instance, this component will check whether a direct object relation presented in Hypothesis is also present in the Text.

The Corpus

For our corpus, we selected a set of 357 iSTART derived Text/SE pairs taken from a recent iSTART experiment. The experiment included 90 high-school students drawn from four 9th grade Biology classes (all taught by the same teacher). The T/SE pairs were assigned by two experts in discourse processing to one of two groups: *Topic identification* (TopicID, n = 96) and *Paraphrase* (n = 261). The major difference between the two main categories was that the TopicID responses tended to include what the sentence was about. Thus, sentences often began with frozen expressions such as "The sentence talks about …". Paraphrase responses, on the other hand, were restatements of the Text, incorporating different words and syntax while lacking any kind of frozen expressions. The Paraphrase group in this study was further subdivided into three sub-category paraphrase types: *Paraphrase Inaccurate* (P-Inaccurate, n = 210); *Paraphrase accurate but Close* (P-Close; n = 16); and *Paraphrase accurate and Distant* (P-Distant, n = 35). P-Inaccurate sentences were defined as a failed paraphrase. For example, a participant may have used similar words to the target sentence but created a sentence with a different meaning.

Table 1: An iSTART text together with participant examples of all four response types.

Text	Sometimes a dark spot can be seen inside the nucleus.
TopicID	yes i know that can be a dartkn spot on .think aboyt what thje sentence
Paraphrase-Inaccurate	in dark spots you can see inside the nucleus and the cell
Paraphrase-Close	if you ever notice that a dark spot can be seen inside the nucleus sometime
Paraphrase-Distant	the nucleus have a dark spot that sometimes be seen.its located in the inside of the nucleus.

P-Close sentences were defined as highly similar to the original sentence in terms of sentence structure and/or content words. P-Distant sentences were defined as highly similar to the original sentence in terms of semantics but different in terms of structure and/or content words. (For an extensive review of the classifications, see Best, Ozuru, & McNamara, 2004; for examples of TopicID and Paraphrase categories, see Table 1 above)

Predictions

We predicted that The Entailer Index would result in a more accurate distinction of the two self-explanation user input types, mainly because of the syntactic and negation handling components. However, because the P-Close sub-category demonstrates very similar lexicon and syntax to the TopicID category (as the name suggests), we predicted weaker results for this distinction. Similarly, the P-Inaccurate subcategory provides self-explanations that typically contain lexical items least like the Text and, consequently, least like TopicID sentences. Thus, we predicted the strongest distinction for this sub-category.

Results

To distinguish the two SE groups, we conducted a discriminant analysis, using the TopicID/Paraphrase categories as the dependent variable. To assess which of the available independent variables (i.e., the Entailment Index, LSA, and the five alternative approaches outlined above) best predicted group membership, the 357 item data set was randomly divided into a training set (67%) and a test set (33%). We conducted an analysis of variance (ANOVA) on the training set data to eliminate any of the seven indices that failed to discriminate between the two groups at $p > .100$. The ANOVA resulted in the lemma index being dropped from the analysis; LSA was retained although its discrimination value was significant only in a 1-tailed test. The ANOVA showed that six variables distinguished the two sentence-type groups; however, because a discriminant analysis is sensitive to collinearity, we followed similar previous studies (e.g., McCarthy et al. 2006) and rejected any variables with a correlation at $r >= .70$, retaining the variables with the larger univariate F-value. This process reduced the indices in the analysis to four: The Entailment Index, Synonymy Index, Word Overlap (2), and LSA.

A discriminant analysis was conducted on the training set and the accuracy of the generated predictions was assessed against the test set. The effect of category for each of the predictor variables (see Table 2) indicates that The Entailment Index was the best predictor of topic identification type self-explanations, $F(1.228) = 25.051$, $p < .001$. The weakest predictor was LSA, $F(1.228) = 2.975$, $p = .086$. The value of the discriminant analysis generated function was significant ($X^2 = 31.18$, $df = 4$, $p < .001$).

Table 2: Effect of category for each predictor variable

	Means		
	TopicID	Paraphrase	F
Entailer	0.60 (0.27)	0.44 (0.20)	25.05**
Synonymy	0.00 (0.02)	0.03 (0.05)	8.64*
Word	0.52 (0.25)	0.44 (0.25)	4.10*
LSA	0.67 (0.35)	0.59 (0.33)	2.98

Note: df = 1,228, SD in parentheses, ** $p < .001$; * $p < .01$

The Fisher's Function Coefficients (see Table 3) demonstrates the direction of the indices used in this analysis. The Entailment Index coefficient values are higher for the TopicID function than for the Paraphrase function. This suggests that self explanations that are subsumed by the Text are more likely to be viewed as identifying the topic rather than a paraphrase of the Text. In contrast, the synonymy index is higher for the Paraphrase category. The synonymy values suggest that synonymous terms are more common to paraphrased responses than to those which identify the topic when other variables have been taken into consideration. The word overlap index is also higher for the Paraphrase category. This result suggests that paraphrased self explanations share more lexical units with their corresponding Text than do TopicID sentences when other variables in the analysis have been taken into consideration. The LSA coefficients are very similar for both categories. This weak distinction suggests that LSA is not a strong discriminator of the categories once other predictors have been taken into consideration.

The accuracy of the discriminant function can best be judged by assessing its generated predictions of category membership against the test set data. The distinction between the two groups was significant ($X^2 = 17.27$, $df = 1$, $p < .001$); however, the accuracy of the predictions was higher for the Paraphrase category: TopicID category (recall = .692; precision = .409); Paraphrase category (recall =

.743; precision = .904). Although the significant results are encouraging, the low precision score for the TopicID category required further analysis.

Table 3: Fisher's unstandardized coefficients for topic identification (TopicID) and paraphrase

	TopicID	Paraphrase
Entailer	9.478	6.009
Synonymy	5.741	16.519
Word Overlap	1.255	1.657
LSA	2.850	2.951
Constant	-4.816	-3.403

Post Hoc Analysis

Precision values are calculated as hits/hits + false alarms. Thus, the low precision value for the TopicID category was caused by a great many false alarms. More simply put, many Paraphrase type SEs were classified as TopicID. The question however, was which of the three Paraphrase sub-categories was most responsible for this problem. To answer this question, we assessed the generated predictions against each of the three Paraphrase sub-categories and found that the sub-category of Paraphrase Close produced the lowest prediction accuracy (Close = 56.35%; Distant = 63.86%; Inaccurate = 74.29%). These results were in line with our predictions, indicating that the more *in*accurate a paraphrase is judged to be, the better it can be distinguished from the topic identification category.

Our next question was which of the indices in the analysis (if any) was contributing to the inaccuracy of evaluating the categories. An individual analysis of the contribution of The Entailer Index and the LSA index revealed that the Entailer index tended to generate higher values for the TopicID group (M = .611, SD = .266) than for the Paraphrase category as a whole (M = .421, SD = 187), the means of the Paraphrase Close sub-category value resembling the Paraphrase category as a whole (M = .482, SD = .234). This direction towards higher values for TopicID was also found for LSA values: TopicID group (M = .679, SD = .334); Paraphrase group as a whole (M = .591, SD = .334). And once more, the means of the Paraphrase Close sub-category value resembled the Paraphrase category as a whole (M = .557, SD = .324). However, this trend of higher values for the TopicID category was reversed for LSA evaluations of *misclassified* items in the Paraphrase Close sub-category (i.e., the sub-category that was least accurate in the analysis). Specifically, considering only the misclassified SEs, the Entailer Index values followed the general trend of higher values for TopicID (TopicID: M = .702, SD = .124; Paraphrase = M = .311, SD = 127). However, the *LSA* values showed the opposite trend, with higher values for Paraphrase (TopicID: M = .504, SD = .344; Paraphrase: M = .599, SD = .322).

Thus, our final question was why this reversal might have occurred. We hypothesized that this reversal might be explained by the text length confound affecting the LSA index. (As described in the introduction, this confound posits that longer texts tend to generate higher LSA values). Thus, the higher LSA means may largely have been caused by longer sentences. And indeed, assessing the lengths of the SEs in the Paraphrase Close sub-category, we can report that the *in*correctly classified SEs were typically longer than the correctly classified SEs (TopicID: M = 17.143, SD = 8.375; Paraphrase: M = 23.778, SD = 7.412). Although the difference in length was not significant: ($F(1,14) = 2.820$, $P = .115$), the effect size ($\eta = .168$) indicates that the difference in sentence length is substantial enough to have affected the analysis. Thus, we conclude that the main cause for the misclassification leading to the low precision results may be attributable to the LSA text length confound.

Discussion

In this study, we assessed the capacity of a variety of computational indices to distinguish two highly similar text types: Topic identification sentences and Paraphrases. The purpose of the study was to offer approaches to improving assessment algorithms of user inputs for Intelligent Tutoring Systems such as iSTART. When incorporated into such systems, the algorithms can be used to provide more accurate and more appropriate feedback to users. Accurate and appropriate feedback facilitates learning and is therefore critical to ITS operating within natural language dialogue.

The results of our study suggested that The Entailment Index in conjunction with the synonymy index, word-overlap, and LSA, significantly distinguished the two self explanation categories under analysis. However, despite this significant result, approximately 44% of one sub-category's sentence pairs (Paraphrases Close) were misclassified. At least part of the cause of this misclassification appears to be the text length confound that affects the LSA index.

The contribution of The Entailment Index to the discriminant algorithm was larger than any other index. Such a result is encouraging for the Entailer. Thus, future development for The Entailer will focus on improving accuracy still further. One approach to this improvement is through changing the way syntactic information is gathered for students' self-explanations. For our experiments in this paper, the syntactic information was gathered from syntactic parsers trained on English sentences written by professional journalists. However, student self explanations are much less grammatically correct than journalists' articles, typically containing textual chunks that are syntactically uncommon. When a parser trained on edited sentences is applied to less correct sentences the retrieved syntactic information is not entirely reliable. This observation leads us to believe that we can employ a partial parser that detects major phrases (Noun Phrases, Verb Phrases) in a sentence without grouping these phrases into a full S (sentence)

structure. The parser will be able to give us syntactic structures for the correct chunks and we can augment this information with an approach using a set of heuristics to be applied to the most frequent *less common* structures used by students. For instance, in the SE given in the introduction *the nucleus the center of the cell it contains the ribsome and more* there is a noun phrase, *the nucleus*, followed by another noun phrase, *the center of the cell*, followed by a pronoun, *it*. Such a sequence is grammatically unlikely and would confuse a full parser. A partial parser, however, would be able to detect the noun phrases, *the nucleus* and *the center of the cell*, along with the pronoun, *it*, but it would not try to group them together into a sentence structure. Thus, we plan to enhance the partial parser with heuristics that would create the missing syntactic relations. For instance, two consecutive articulated noun phrases would be separated by a linking verb. Similarly, a noun phrase followed by a pronoun followed by a verb would indicate that the pronoun starts a new sentence. We predict that these two heuristics would substantially increase the likelihood of solving the incorrect structures of typical student input.

The application of findings in cognitive science to Intelligent Tutoring Systems is largely dependent upon the accuracy of the algorithms that assess the systems' input. This study demonstrates a new and successful approach to assessing such input. Future research will focus on developing this accuracy still further so as to offer systems such as iSTART the algorithms necessary to provide learners with the most appropriate feedback. Accurate feedback is a critical aspect of learning and the development of the Entailment Index may play a critical role in enhancing that accuracy.

Acknowledgements

This research was supported by the Institute for Education Sciences (IES; R305G020018, R305G040046) and partially by the National Science Foundation (NSF; REC-0241144). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES or NSF.

References

Aleven, V., & Koedinger, K. R. (2002). An Effective Meta-cognitive Strategy: Learning by Doing and Explaining with a Computer-Based Cognitive Tutor. *Cognitive Science, 26*, 147-179.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1989). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4*, 167-207.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*. 111-127.

Best, R., Ozuru, Y., & McNamara, D. S. (2004). Self-explaining science texts: Strategies, knowledge, and reading skill. In Y. B. Kafai, W. A. Sandoval, N. Enyedy, A. S. Nixon, & F. Herrera (Eds.), p*roceedings of the Sixth International Conference of the Learning Sciences: Embracing diversity in the learning sciences* (pp. 89-96). Mahwah, NJ: Erlbaum.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of North American Chapter of Association for Computational Linguistics (NAACL-2000)*.

Dagan, I., Glickman, O., & Magnini, B. (2005). The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Recognizing Textual Entailment Challenge Workshop*.

Kluger, A. N., & DeLisi, A. (1996). The effects of feedback interventions on performance: A historical review, a metaanalysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119, 254-284.

Graesser, A.C., Chipman, P., Haynes, B.C., and Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education* 48, 612-618.

Landauer, T., McNamara, D. S., Dennis, S., and Kintsch, W. Eds. (2006). The Handbook of *LSA.* Mahwah, NJ: Erlbaum.

Magerman, D. (1994). *Natural Language Parsing as Statistical Pattern Recognition*. Ph.D. Dissertation, Stanford University.

McCarthy, P.M, Lewis, G.A., Dufty, D.F., & McNamara, D.S. (2006). Analyzing writing styles with Coh-Metrix. In *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS)*, Melbourne, Florida.

McCarthy, P.M., Rus, V., Crossley, S.A., Bigham, S.C., Graesser, A.C., & McNamara, D.S. (In Press). Assessing the *Entailer* with a Corpus of Natural Language From an Intelligent Tutoring System. Submitted to FLAIRS, 2007.

McKendree, J. (1990). Effective feedback content for tutoring complex skills. *Human-Computer Interaction*, 5, 381-413.

McNamara, D.; Levinstein, I. B.; & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, and Computers* 36:222–233.

Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM* 38.39–41.

Rus, V., Graesser, A.C., McCarthy, P. &Lin, K. (2005). A lexico-syntactic approach to textual entailment. In *Proceedings of International Conference on Tools with Artificial Intelligence.*

Rus, V., McCarthy, P.M., & Graesser, A.C. (2006). Analysis of a Textual Entailer. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing-06), February, Mexico City, Mexico. Lecture Notes in Computer Science, Vol. 3878, Springer.

Sims-Knight, J.E., Upchurch, R.L. (2001). What's Wrong with Giving Students Feedback? Proceedings of the 2001 ASEE Annual Conference, Albuquerque, NM.