Proceedings of the Twenty-First International FLAIRS Conference (2008)

Assessing Forward-, Reverse-, and Average-Entailer Indices on Natural Language Input from the Intelligent Tutoring System, iSTART

Philip M. McCarthy, Vasile Rus, Scott A. Crossley*, Arthur C. Graesser, & Danielle S. McNamara

Institute for Intelligent Systems
University of Memphis
Memphis. TN 38152
{pmmccrth, vrus, a-graesser, dsmcnamr} @ memphis.edu; *scrossley@mail.psyc.memphis.edu

Abstract

This study reports on an experiment that analyzes a variety of entailment evaluations provided by a lexico-syntactic tool, the *Entailer*. The environment for these analyses is from a corpus of self-explanations taken from the Intelligent Tutoring System, iSTART. The purpose of this study is to examine how evaluations of hand-coded *entailment*, *paraphrase*, and *elaboration* compare to various evaluations provided by the Entailer. The evaluations include standard entailment (forward) as well as the new indices of Reverse- and Average-Entailment. The study finds that the Entailer's indices match or surpass human evaluators in making textual evaluations. The findings have important implications for providing accurate and appropriate feedback to users of Intelligent Tutoring Systems.

Introduction

A major challenge for Intelligent Tutoring Systems (ITSs) that incorporate natural language interaction is to accurately evaluate users' contributions and to produce appropriate 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 (Mark & Greer, 1995; McKendree, 1990). The benefits of feedback in ITSs are equally evident (Azevedo & Bernard, 1995). This study addresses the challenge of evaluating users' textual input in ITS environments by reporting the results of an experiment conducted on data from the ITS, iSTART (Interactive Strategy Training for Active Reading and Thinking; McNamara, Levinstein, & Boonthum, 2004). More specifically, we concentrate on a variety of *entailment* evaluations that are generated from a lexico-syntactic computational tool, the *Entailer* (Rus et al., in press [a], [b]). The results of the experiment demonstrate that evaluations generated from the Entailer present a viable and accurate approach to assessing natural language input in ITS environments and, through these evaluations, the capacity to present appropriate feedback to ITS users.

The ITS in this study, iSTART, uses pedagogical agents to provide young adolescent to college-aged students with tutored self-explanation and reading strategy training. iSTART is designed to improve students' ability to self-explain by teaching them to use reading strategies such as *elaboration*, *bridging*, and *paraphrasing*. For example, paraphrasing requires students to restate sentences in their own words; such a process helps students to monitor their comprehension and also to activate knowledge relevant to the target information. Following the *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.*

The object of existing iSTART algorithms is to assess which strategy (or, *type* of self-explanation) has been attempted by the student. However, further algorithms are needed to assess how close in meaning the self-explanation is to the target sentence (i.e., is the self-explanation a *paraphrase* of the target sentence? Is it an *elaboration*? Or is it *entailed* by the target sentence?). Thus, the more accurately the self-explanations can be assessed, the more appropriately the system can provide feedback to the user. In this study, we explore these evaluations of self-explanations using a variety of textual-assessment metrics.

Computationally Assessing Text Relatedness

Established text relatedness metrics such as Latent Semantic Analysis (LSA; Landauer et al., 2007) and overlap indices have proven to be extremely effective measures for a great variety of the systems that analyze natural language and discourse, such as Coh-Metrix (Graesser et al., 2004), iSTART (McNamara et al., 2004), and AutoTutor (Graesser et al., 2005). Despite such successes, the need remains for new measures of textual assessment to augment existing measures and thereby better assess textual comparisons. In this study, we assess a variety of more established textual relatedness assessment metrics (e.g., LSA and Content-Overlap), and compare them to newer approaches such as the Entailer (Rus et al., in press [a]) and *MED* (McCarthy et al., 2007). Each of these measures provides a unique approach to assessing the relatedness between text fragments. Therefore, we believe that a combination of such approaches is likely to offer the user the most complete range of feedback.

Latent Semantic Analysis. LSA is a statistical technique for representing world knowledge based on large corpora of texts (Landauer et al., 2007). LSA uses a general form of factor analysis (singular value decomposition) to condense a high dimensional representation of a very large corpus of texts to 300-500 dimensions. These dimensions represent how words (or group of words) co-occur across a range of documents within a large corpus (or space). Unlike content overlap indices, LSA affords tracking words that are semantically similar, even when they may differ morphologically.

Content Overlap. Content-overlap assesses how often a common noun/verb/adjective co-occurs in adjacent sentences. Such measures have been shown to aid in text comprehension and reading speed (Kintsch & Van Dijk, 1978).

Minimal Edit Distances (MED). *MED* (McCarthy et al., 2007) is a computational tool designed to evaluate text relatedness by assessing the similarity of the lexical items in adjacent sentences. *MED* is a combination of measuring Levenstein distances (1966) and string theory matching (Dennis, 2006). Essentially, *MED* functions like a spellchecker; that is, it converts words into unique character representations and then searches for the shortest route through which two such characters can be matched. The evaluations of the possible routes result in a set of *costs*: shifting the string (right or left) has a cost of 1; deleting a character costs 1; and inserting a character costs 1. Once the cost has been calculated, the value is divided by the number of elements in the string. *MED* scores are continuous, with a score of zero representing an identical match. *MED*'s major benefit over simple co-occurrence indices is that structural variation can be assessed. Thus, for *MED*, *the cat chased the dog* is different from *the dog chased the cat* (see Table 1).

Table 1. *MED* Evaluations for "The dog chased the cat."

	MED
The dog chased the cat.	0
The cat chased the dog.	0.267
The cats chased the dogs.	0.533
The cat didn't chase the dog.	0.941
Elephants tend to be larger than mice.	1.263

Entailer. The purpose of the Entailer is to evaluate the degree to which one text is entailed by another text. We say that T (the entailing text) entails H (the entailed hypothesis). The Entailer is based on the industry approved testing ground of the *recognizing textual entailment* (RTE) corpus (Dagan, Glickman, & Magnini, 2004-2005). The Entailer uses minimal knowledge resources and delivers high performance results compared to similar systems (Rus et al., in press [a], [b]).

The approach adopted by the Entailer encompasses lexico-syntactic information, negation handling, synonymy, and antonymy. The Entailer addresses two forms of negation: *explicit* and *implicit*. Explicit negation is indicated in the text through surface clues such as *n't*, *not*, *neither*, and *nor*. Implicit negation incorporates antonymy relations between words as encoded in WordNet (Miller, 1995). In our negation score, we compute how many negation markers there are in the Text and Hypothesis. The Text and Hypothesis are required to have the same negation parity in order to have same polarity.

The Entailer functions by having each pair of text fragments (assigned as text [T] and hypothesis [H]) mapped into two graphs, one for T and one for H, with nodes representing main concepts and links indicating syntactic dependencies among concepts as encoded in T and H, respectively. An entailment score, entscore (T,H), is then computed quantifying the degree to which the T-graph subsumes the H-graph. The score is the weighted sum of one lexical and one syntactic component. The lexical component expresses the degree of subsumption between H and T at word level (i.e. vertex-level), while the syntactic component operates similarly at syntactic-relationship level (i.e., edge-level). The weights of lexical and syntactic matching are given by the parameters α and β, respectively (see Equation 1). The effect of negation on the entailment decision is captured by the last term of the equation. An odd number of negation relations between T and H, denoted *#neg_rel*, leads to an entailment score of 0, whereas an even number does not change the bare lexico-syntactic score. The choice of α, β and γ can have a large impact on the overall score. From its definition, the entailment score is non-reflexive, entscore(H, T) ≠ entscore(T,H), because it is normalized based on the characteristics of the hypothesis ($|V_h|$ and $|E_h|$). Thus, if one reverses the roles of T and H, the normalizing factor will change (see Rus et al., in press, for a full discussion).

Equation 1. Scoring formula for graph subsumption.

$$entscore(T,H) = (\alpha \times \frac{\sum_{V_h \in H_v} \max_{V_t \in T_v} match(V_h, V_t)}{|V_h|} + \beta \times \frac{\sum_{E_h \in H_e} \max_{E_t \in T_e} match(E_h, E_t)}{|E_h|} + \gamma) \times (\frac{1+(-1)^{\#neg\_rel}}{2})$$

In Equation 1, α, β, and γ are in the [0,1] interval. γ is a free term, which can be used to bias the entailer (e.g., to give higher or lower scores). Thus, if an *optimistic* entailer is required, γ can be set to a high value. The values for α, β, and γ should always add up to 1.0.

For the purposes of natural language assessment in ITSs, the Entailer offers many advantages over current text relatedness measures such as LSA. First, because lexical/word information acts only as a component of the overall formula, the Entailer is less susceptible to the text length confound (i.e., longer sentence pairs tending to generate higher values, see McCarthy et al., in press). Because the Entailer addresses both syntactical relations and negation, the tendency for higher relatedness results over lengthier texts is reduced. Also, the Entailer addresses asymmetrical issues by evaluating text non-symetrically, so entscore(H, T) ≠ entscore(T,H). Thus, the evaluation of a response (self explanation) to a stimulus (target text) will be different from the evaluation of the stimulus to the response. Finally, the Entailer handles negations so it offers the opportunity of providing more accurate feedback.

The effectiveness of the Entailer has been assessed and validated over numerous studies. These studies include comparisons to similar systems (Rus et al., in press [a], [b]), evaluations of negation and synonyms components (Rus et al., in press [a], [b]), assessments of entailment evaluations on the standard RTE data set (Rus et al., in press [a], [b]) and Microsoft Research Paraphrase Corpus (Dolan, Quirk, & Brockett, 2004; Rus, McCarthy, & Graesser, in press), and evaluations of the Entailer using natural language input from ITSs (McCarthy et al., 2007; Rus et al., in press [b]).

Reverse- and Average-Entailment

In the context of an ITS, the text [T] typically represents the *target sentence* and the hypothesis [H] typically represents the student input. Thus, in Rus et al. (in press [b]), when the Entailer was applied to AutoTutor (Graesser et al., 2005), the target sentence was a stored *ideal answer*. In the context of iSTART (McNamara et al., 2004), the target sentence is presented to the student user as a sentence that is required to be self-explained. Meanwhile, the hypothesis fragment [H] is the fragment assumed to be entailed by the text [T]. In both AutoTutor and iSTART environments, H is the user input.

To date, such a T/H assumption has proven effective, with the Entailer's results consistently outperforming competing textual analysis metrics (see Rus et al., in press [b]). However, in this study, we expand the assessment of the Entailer by considering two additional indices. Thus, the main index for the Entailer we now call *forward Entailment* (*Ent-for*). This is the entailment index that has been used on all previous entailment experiments (e.g. Rus et al., in press [a], [b]). Specifically, *Ent-for* is normalized by dividing the number of paired matches by the number of nodes in H (i.e., the hypothesis sentence). The second entailment index we consider in this study we label *Reverse-entailment* (*Ent-rev*). The initial calculation of *Ent-rev* is the same as *Ent-for*; however, for *Ent-rev* the division is based on the number of nodes in T (i.e. the target sentence). That is, *Ent-rev* assumes that the target sentence is entailed by the student input. The index is useful because, in the iSTART context, student *elaborative* responses may be longer than the target sentence. In such cases, the response may entail the target and, therefore, appropriate feedback needs to be supplied to the user.

In this study, we also assess *Average-entailment* (*Ent-avg*; Rus et al., in press). *Ent-avg* is the mean of the forward and reverse evaluations. *Ent-avg* is introduced to better assess potential *paraphrased* student input. In paraphrase assessment, the target sentence can be assumed to entail the student response to a similar degree as the response entails the target. Thus, *ent-avg* is predicted to be a better assessment of paraphrase than the two alternative entailment indices.

Corpus

In order to test the textual relatedness approaches outlined above, we used a natural language corpus of iSTART user input statements. The data *pairs* used to make the corpus were generated from an iSTART experiment conducted with 90 high-school students drawn from four 9th grade Biology classes (all taught by the same teacher). Overall, the experiment generated 826 sentence pairs. The average length of the combined sentence pairs was 16.65 words (*SD* = 5.63). As an example, the following four self-explanations were made (reproduced without correction) by students responding to the target sentence *Sometimes a dark spot can be seen inside the nucleus*:

1) yes i know that can be a dartkn spot on .think aboyt what thje sentence
2) in dark spots you can see inside the nucleus and the cell
3) if you ever notice that a dark spot can be seen inside the nucleus sometime
4) the nucleus have a dark spot that sometimes be seen.its located in the inside of the nucleus.

To assess the pairs, three discourse processing experts evaluated each sentence pair on each of the three dimensions of similarity: paraphrase, entailment, elaboration (see Table 3 for examples). We use the term entailment to refer to *explicit* textual reference. As such, we distinguish it from the term *implicature* for which references are only *implied* (see McCarthy et al., 2007). We use the term *elaboration* to refer to any student input information that is generated as a *response* to the stimulus text without being a case of entailment or implicature. An elaboration may differ markedly from its textual pair provided it does not contradict either the target text or world knowledge. In such an event, the input would be considered as simply an error. We use the term *paraphrase* to mean a reasonable restatement of the text. Thus, a paraphrase tends to be an entailment, but an entailment does not have to be a paraphrase. For example, the sentence *the dog has teeth* is entailed by (but not a paraphrase of) the sentence *the dog bit the man*. Because a paraphrase must also be an entailment, evaluations of the two textual similarity dimensions will tend to correlate. Similarly, responses that are elaborations will negatively correlate with both paraphrase and entailment measures. Thus, distinguishing the textual similarity measures in order to provide optimal feedback is not a trivial task.

Table 3. Categorization of Responses for the Target Sentence *John drove to the store to buy supplies.*

Category	Student Statement	Relationship to Source Sentence
Entailment	John went to the store.	Explicit, logical implication
Elaboration	He could have borrowed stuff.	Non-contradictory reaction
Paraphrase	He took his car to the store to get things that he wanted.	Reasonable restatement

The three raters were first trained on a subset (100 data pairs) of the *iSTART* corpus. Each pair (for each category) was given a rating of 1 (min) to 6 (max). Training was considered successful when the *r* value for each of the three categories was .75 or above. For the final analysis, a Pearson correlation for each inference type was conducted between all possible pairs of raters' responses using all the available pairs. Correlation evaluations for agreement not only provide a measure of typical human evaluation, they also serve to evaluate the computational model in comparison to such experts. In addition, because the output of the Entailer is a continuous value, correlations are a practical and effective evaluation of the efficacy of the system.

As Hatch and Lazarton (1991) point out, a greater number of raters increases the confidence of the inter-rater reliability. As such, we follow the Hatch and Lazaraton formula to convert multiple raters' correlations into a single effective gold inter-rater value. Thus, the effective inter-rater reliability for the Pearson correlations were as follows: paraphrase (r = .909), entailment (r = .846), elaboration (r = .819). The *individual* correlations and their averages are presented in Table 4.

Table 4: Inter-rater correlations and average for three categories of textual similarity

Raters	Paraphrase	Entailment	Elaboration
1-2	0.720	0.595	0.630
1-3	0.793	0.685	0.609
2-3	0.771	0.688	0.604
Average	0.761	0.656	0.614

For this experiment, we removed from the corpus two types of pairings that did not suit this analysis. First, we removed any pair in which the response was noticeably garbage (e.g., where the user had randomly hit the keyboard or where responses consisted of no more than one word; such responses would be filtered out of the iSTART system under normal operation). Second, we removed pairs where the target sentence was a question. For these pairs, users had tried to answer the question rather than self-explain it. Following these removals, our final corpus consisted of 631 pairs. This corpus was further divided into two groups for training (419 cases) and testing (212 cases) purposes.

Predictions

Following previous Entailer comparison studies (e.g., McCarthy et al., 2007), we predicted that Entailer would outperform other textual comparison measures such as LSA. Specifically, for *human paraphrase* evaluations, where T entails H to the same degree that H entails T, we predicted *Ent-avg* to produce the greatest accuracy of evaluation. For *human entailment* evaluations, where T entails H, we predicted that *Ent-for* would produce the highest accuracy of evaluations. And for *human elaboration* evaluations, where H tends to be longer than T, we predicted *Ent-rev* to be the most accurate index.

Results

Correlations

The correlation results (based on the *training* data) largely confirmed our predictions (see Table 5). The Entailer's indices produced the highest correlations with human evaluations (paraphrase: *r*=.818, *p*<.001; entailment: *r*=.741, *p*<.001: elaboration: *r*=-.673, *p* <.001). Comparing the Content-overlap and LSA indices, the former was significantly more accurate (paraphrase: *z-diff* = 1.984, *p*=.047; entailment: *z-diff* = 2.000, *p*=.045; elaboration: *z-diff*=1.827, *p*=.068). There was no significant difference between evaluations of LSA, Content-overlap, and *MED*.

Of the three Entailer indices, the highest correlating index for paraphrase was *Ent-rev* (*r*=.818, *p*<.001). There was no significant difference between this value and that of *Ent-avg* (*r*= .769, *p*<.001), which was our predicted index. We speculate that the apparently higher *Ent-rev* value results from student responses for paraphrase being longer than their corresponding target sentence. Given that a target sentence could be considered the *ideal form* of the sentence, a student trying to paraphrase that sentence would probably have to use more words, which indeed they appear to have done when length is compared (*F* (1, 1334) = 28.686, *p*<.001).

Table 5: Correlations between comparison type and text evaluation measure (n = 419)

Paraphrase	Ent-Rev	Ent-Avg	Content	MED	LSA	Ent-For
	0.818	0.769	0.659	0.634	0.574	0.566
Entailment	Ent-For	Ent-Avg	MED	Ent-Rev	Content	LSA
	0.741	0.724	0.578	0.577	0.57	0.469
Elaboration	Ent-For	Ent-Avg	Content	MED	LSA	Ent-Rev
	-0.673	-0.576	-0.515	-0.443	-0.416	-0.380

Note: All correlations are significant at p. < .001.

Regression

Our analysis of the *three* human-coded text relatedness evaluations (paraphrases, entailment, elaboration) consisted of a series of *forced entry* linear regressions, selected as a conservative form of multivariate analysis. Regression was selected as the method of analysis because the dependent variables are continuous. One advantage of regression analysis is that derived values generated from b-weights offer a continuous evaluation of each assessment (in this case, 1-6).

Ultimately, parameters using this scale can be used to assess optimal ranges that most accurately assess the kind of student input (i.e., an *entailed*, *paraphrased*, or *elaborative* response). The hand coded evaluations of *entailment*, *elaboration*, and *paraphrase* were the dependent variables and the computational evaluation index with the highest correlation to the training set data were used as independent variables. The results below are based on the *test* set data using the coefficients derived from the regressions on the *training* set data.

Paraphrase. Using *Ent-rev* as the independent variable, a significant model emerged, *F* (1, 417) = 844.151, *p* < .001. The model explained 66.9% of the variance (Adjusted R^2 = .669). *Ent-rev* was a significant predictor (*t* = 29.054, *p* < .001). The derived B-weights were then used to calculate the accuracy of the model against the held-back *test set* data (*n*=212). The correlation between the derived evaluation and the hand-coded paraphrase values was high (*r*=.840, *p*<.001). Indeed, the correlation was significantly higher than that produced by the mean of the human coders (*z-diff* = 2.874, *p* =.004), suggesting that the model is at least as accurate as the expert raters. When the other indices were added to the model there was no significant increase in accuracy. Replacing *Ent-rev* with *Ent-avg* resulted in significantly lower accuracy (*r* =.755, *p*<.001; *z-diff* = 2.420, *p* = .016).

Entailment. Using *Ent-for* as the independent variable, a significant model emerged, *F* (1, 417) = 507.936, *p*< .001. The model explained 54.8% of the variance (Adjusted R^2 = .548). *Ent-for* was a significant predictor (*t* = 22.537, *p* < .001). The derived b-weights were then used to calculate the accuracy of the model against the held-back *test set* data (*n*=212). The correlation between the derived evaluation and the hand-coded entailment values was high (*r*=.708, *p* <.001); the correlation was not significantly different from that produced by the human coders, suggesting that the model is at least as accurate as the three expert raters. When the other indices were added to the model there was no significant increase in accuracy.

Elaboration. Using *Ent-for* as the independent variable, a significant model emerged, *F* (1, 417) = 345.715, *p* < .001. The model explained 45.2% of the variance (Adjusted R^2 = .452). *Ent-for* was a significant predictor (*t* = -18.593, *p* < .001). The derived b-weights were then used to calculate the accuracy of the model against the held-back *test set* data (*n*=212). The correlation between the derived evaluation and the hand-coded entailment values was again high (*r*=.676, *p*<.001); the correlation was not significantly different from that produced by the human coders, suggesting that the model is at least as accurate as the three expert raters. When the other indices were added to the model there was no significant increase in accuracy.

Discussion

In this study, we compared various indices derived from the Entailer, a computational tool that evaluates the degree to which one text is entailed by another, to a variety of other text relatedness metrics (e.g., LSA). Our corpus was formed from 631 iSTART target-sentence/self-explanation pairs. The self-explanations were hand coded across three categories of text relatedness: *paraphrase*, *entailment*, and *elaboration*. A series of regression analyses suggested that the *Entailer* was the best measure for approximating these hand coded values. The *Ent-rev* index of the *Entailer* explained approximately 67% of the variance for paraphrase; the *Ent-for* index explained approximately 55% of the variance for entailment, and 45% of the variance for elaboration. For each model, the derived evaluations either met or surpassed human inter-rater correlations, meaning that the algorithms can produce assessments of text at least equal to that of three experts.

The accuracy of our models is highly encouraging. Future work will now move towards implementing algorithms that use these Entailer evaluations to provide feedback to students and assess that feedback when applied to users of the iSTART system. As each model produces a value between approximately 1 and 6, we envision that dividing these values into *low* (e.g. <2.67), *moderate* (e.g. >2.67 < 4.33), and *high* (e.g. >4.33) partitions will allow us to provide users with accurate feedback on their input. For example, a moderate paraphrase evaluation, coupled with a high elaboration evaluation might call for a feedback response such as "*Your paraphrase is fairly good. However, you have included a lot of information that is not* *really relevant. See if you can write your paraphrase again with more information from the target sentence, and reduce the information that is not in the target sentence.*"

While only the Entailer indices contributed to the final assessment models, all other measures (e.g., LSA) correlated highly with the hand coded evaluations. This is important because these other measures are still envisioned to contribute to a final feedback algorithm. Specifically, a high content-overlap evaluation coupled with a high paraphrase evaluation could indicate that a paraphrase may have been successful only because many of the words from the target sentence were reproduced in the response.

This study builds on the recent major developments in assessing text relatedness indices, particularly the incorporation of strings of indices designed to assess natural language input in ITSs. Research has shown that the Entailer delivers high performance analyses when compared to similar systems in the industry approved testing ground of *Recognizing Textual Entailment* tasks. However, the natural language input from the corpus in this study (with its spelling, grammar, and syntax issues) provided a far sterner test in which the performance of the Entailer has been significantly better than comparable approaches. This finding is compelling because accurate assessment metrics are necessary to better evaluate input and supply optimal feedback to students. This study offers promising developments in this endeavor.

Acknowledgements

This research was supported by the Institute for Education Sciences (IES R305G020018-02) and partially by the National Science Foundation (REC 0106965 and ITR 0325428). The authors thank Rachel Best, Sarah Bigham, Stephen Briner, Erin Lightman, Yasuhiro Ozuru, and Adam Renner for their contribution.

References

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

Dagan, I., Glickman, O., and Magnini, B. 2004 - 2005. Recognizing textual entailment. *Pattern Analysis, Statistical Modeling and Computational Learning*.

Dennis, S. 2006. Introducing word order in an LSA framework. In *Handbook of Latent Semantic Analysis*. T. Landauer, D. McNamara, S. Dennis and W. Kintsch eds.: Erlbaum.

Dolan, W. B., Quirk, C. & Brockett, C. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of COLING 2004*.

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.

Graesser, A., McNamara, D., Louwerse, M., & Cai, Z. 2004. *Coh-Metrix*: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers* 36, 193-202.

Hatch, E. & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* Boston, MA: Heinle & Heinle.

Kintsch, W. and Van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363-394

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. 2007. *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

Levenstein, V. (1966). Binary codes capable of correcting insertions and reversals. Soviet. Phys. Dokl., 10, 707-717.

Mark, M.A & Greer, J.E. 1995. The VCR Tutor: Effective instruction for device operation. *The Journal of the Learning Sciences 4*: 209-246.

McCarthy, P.M., Renner, A. M. Duncan, M.G., Duran, N.D., Lightman, E.J., & McNamara. D.S., In press. Identifying topic sentencehood. *Behavior, Research and Methods, Instruments, and Computers.*

McCarthy, P. M., Rus, V., Crossley, S. A., Bigham, S. C., Graesser, A. C. & McNamara, D. S. 2007. Assessing entailer with a corpus of natural language from an intelligent tutoring system. *Proceedings of the 20th International Florida Artificial Intelligence Research Society* (pp. 247-252). Menlo Park, California: AAAI Press.

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

McNamara, D. S., Levinstein, I. B., and Boonthum, C. 2004. iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instrument, & Computers* 36: 222-233.

Miller, G.A. 1995. WordNet: A lexical database for English. In *Communications of the ACM* 38: 39-41.

Rus, V., Graesser, A.C., McCarthy, P.M., and Lin, K. 2005. A Study on Textual Entailment, *IEEE's International Conference on Tools with Artificial Intelligence*. Hong Kong.

Rus, V., McCarthy, P.M., & Graesser, A.C. (in press). Paraphrase Identification with Lexico-Syntactic Graph Subsumption. FLAIRS 2008.

Rus, V., McCarthy, P.M., Lintean, M.C., Graesser, A.C., & McNamara, D.S., (2007). Assessing Student Self-Explanations in an Intelligent Tutoring System. *Proceedings of the 29th annual conference of the Cognitive Science Society*. Cognitive Science Society.

Rus, V., McCarthy, P.M., McNamara, D.S., & Graesser, A.C. (in press [a]). A study of textual entailment. *International Journal on Artificial Intelligence Tools.*

Rus, V., McCarthy, P.M., McNamara, D.S., & Graesser, A.C. (in press [b]). Natural Language Understanding and Assessment. *Encyclopedia of Artificial Intelligence*.