Appear in Proceedings of the International Conference on Artificial Intelligence (IJCAI2007), 2007

A Hybrid Ontology Directed Feedback Selection Algorithm for Supporting Creative Problem Solving Dialogues*

Hao-Chuan Wang[1], Rohit Kumar[1], Carolyn Penstein Rosé[1], Tsai-Yen Li[2], Chun-Yen Chang[3]
[1]Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213
[2]Computer Science Department, National Chengchi University, Taipei, Taiwan
[3]Science Education Center, National Taiwan Normal University, Taipei, Taiwan
{haochuan,rohitk,cprose}@cs.cmu.edu, li@cs.nccu.edu.tw, changcy@ntnu.edu.tw

Abstract

We evaluate a new hybrid language processing approach designed for interactive applications that maintain an interaction with users over multiple turns. Specifically, we describe a method for using a simple topic hierarchy in combination with a standard information retrieval measure of semantic similarity to reason about the selection of appropriate feedback in response to extended language inputs in the context of an interactive tutorial system designed to support creative problem solving. Our evaluation demonstrates the value of using a machine learning approach that takes feedback from experts into account for optimizing the hierarchy based feedback selection strategy.

1 Introduction

In this paper we describe and evaluate a new hybrid language processing approach designed for interactive language processing applications that maintain an extended interaction with users over multiple turns. Except in the case of complete user initiative, the system must have some explicit representation of the state-space of reasonable dialogues in the form of a hierarchy of dialogue recipes or templates [Rosé et al., 2001], or an ontology of dialogue topics [Popescu, Aleven, and Koedinger, 2003] or discourse states [Tetreault & Litman, 2006] in order to ensure coherence at the level of topic transitions and ultimately task success. It is also necessary to ensure robust understanding of user input. While state-of-the-art systems typically achieve robust understanding by encouraging users to contribute concise contributions in order to avoid recognition errors [Litman & Pan, 2002], recently a new emphasis on applications that require more sophisticated language from users, such as tutorial dialogue systems [Wang et al., 2006; Rosé & VanLehn, 2005; Popescu, Aleven, & Koedinger, 2003], is emerging as a new focus area for the computational linguistics community.

A major thrust of research in the language technologies community for the past decade has been the development of shallow language processing approaches that require very little knowledge engineering and that can be applied efficiently to very large corpora. Typically text categorization algorithms and text retrieval algorithms operate on large portions of text, such as whole news articles [Ko & Seo, 2004]. However, while this technology has frequently been applied to dialogue applications where users contribute extended contributions [e.g., Graesser et al., 2001; Malatesta et al., 2002], they have often been met with less success than in more typical text retrieval applications. The demands of interactive language processing systems are different from more typical text retrieval applications because texts are much shorter, precision is more important, and domain specific development data is in short supply. Nevertheless, the performance of shallow language processing approaches can be enhanced by drawing upon the knowledge sources that must be part of the system for the purposes described above.

We describe a method for using a simple topic hierarchy in combination with a standard information retrieval measure of semantic similarity to reason about extended language inputs in the context of an interactive tutorial system designed to support creative problem solving. Our evaluation demonstrates that while a common-sense approach to combining these two sources of information provides some advantage over shallow language processing alone, that a more substantial improvement can be achieved by tuning the integration algorithm using a machine learning approach that takes feedback from experts into account.

2 Motivation for Hierarchy Based Feedback

Supporting creative problem solving of ill-structured problems is a new direction in the Intelligent Tutoring Community [Aleven et al., 2006], which poses specific challenges for language processing technology that can be used to facilitate this type of learning. In this paper we specifically address the problem of eliciting ideation. While

* This work was funded in part by NSF Grant SBE0354420 and the National Science Council [NSC] of Taiwan under contract no. NSC94-2524-S-003-014.

supporting knowledge construction and reflection can be accomplished using dialogue strategies where the system has the initiative, supporting student ideation requires abdicating control to the student. Systems such as the Geometry Explanation Tutor [Popescu, Aleven, & Koedinger, 2003] and Auto-Tutor [Graesser et al., 2001] support this student directed ideation using a reactive rather than a proactive dialogue management strategy.

Our baseline hierarchy based feedback approach [Wang et al., 2006] is similar in spirit to that adopted in the Geometry Explanation Tutor and Auto-Tutor. However, our approach differs from this prior work in several important respects. First, similar to Popescu et al. [2003], we attach feedback messages to nodes in our hierarchy so that we can use a match between a student contribution and a node in the hierarchy as a basis for selecting a feedback message. However, in contrast to Popescu et al. [2003], we do not utilize a deep, symbolic approach to language understanding. Instead, we attach a number of prototype texts to each leaf node in the hierarchy and determine which node to match to based on a shallow semantic similarity measure. In our approach, the value of the similarity score is a key component in our strategy selection process. If the match is deemed good based on our shallow similarity measure, we select the matched node as the basis for our feedback selection. Otherwise, we move up a level of abstraction in the hierarchy in order to compensate for the partial match. Thus, our approach is much lighter weight in that it does not require heavy knowledge representation or inferencing technology.

Similar to Graesser et al., [2001] we make use of a finite state machine to determine how to use the hierarchy to select feedback. However, in contrast to the Auto-Tutor approach, our strategy is motivated more by general principles of dialogue coherence rather than a specific knowledge elicitation strategy designed to elicit a specific idea from a student with progressively more pointed hints.

3 Technical Approach in Detail

Our system accepts student input in Chinese, although our approach can easily be applied to other languages. As a preprocessing step, the Chinese text must be segmented into individual lexical items using technology provided by the Chinese Knowledge and Information Processing (CKIP) group at Academia Sinica [Ma & Chen, 2003]. This preprocessing step is not necessary except for languages like Chinese and Japanese that have no spaces between words. A word vector is then assembled from the tokenized text, with each position in the vector corresponding to a single lexical item, and term weight equal to a function of term frequency within and across instances, referred to as TF-IDF. Note that text classification using simple word vectors such as this can be done effectively without any morphological processing even for highly synthetic languages such as German [Donmez et al., 2005].

A shallow semantic similarity measure is computed between texts by computing the cosine correlation between their respective word vectors. This semantic similarity measure is used to select the best matching idea node in the hierarchy. Each idea node in the hierarchy is associated with a list of prototype texts. The idea node associated with the text that is rated as most similar to the student text is selected as the best matching node. The magnitude of the computed semantic similarity is used to estimate how good the match is.

We make use of two Finite State Machines (FSMs) to guide the strategy for selecting feedback messages based on the best matching node, the estimated goodness of the match, and a record of which idea nodes have been uttered by the student in previous contributions. In our FSM based approach to feedback selection, a finite set of *states*, Q, represents the range of possible system functional behavior, including actions such as checking the coverage of idea nodes attached to an abstract concept node or moving from one node to a higher node in the hierarchy representing a more abstract idea. The finite *alphabet*, Σ, represents the set of events that can trigger a transition from one state to another, such as the condition that all idea nodes subsumed by the current node have been covered, which in this case triggers a transition to a state requiring the system to select a new conceptual node.

The *feedback* produced by the system consists of two parts. The first part is referred to as a *Comment*, which is meant to offer feedback specifically directed at the details of what students have contributed. The comment text associated with the most closely matching idea node is selected, unless the goodness of the match is rated as low, in which case a more abstract node that subsumes that idea node is selected instead. The second portion of the feedback is referred to as a *Tutorial*, the purpose of which is to direct the student to a new idea, preferably which coherently follows from the current idea if such a subsequent focus exists in the hierarchy and has not yet been covered by the student. In this way students have the opportunity to learn how to evaluate their ideas, and they are encouraged to continue contributing additional ideas. Two separate FSMs are designed for the selection of each of the two feedback portions. In the current design, a *Comment* is selected based on a function only of the current student contribution, while a *Tutorial* is selected based on a function of the student contribution and the context of students' previous responses within the same session. This is what we refer to as the Hierarchy approach to feedback selection.

To further illustrate our Hierarchy approach to feedback selection as well as to demonstrate where the value of using a simple idea Hierarchy in feedback selection can be found, we compare our Hierarchy approach to a baseline No Hierarchy approach using the example in Table 1. In the No Hierarchy approach, the Comment is selected based strictly on the best matching idea node, and the tutorial is selected

Table 1 Example of multi-turn interaction contrasting the With Hierarchy approach to feedback selection with the No Hierarchy approach.

Student Response	Matched Concept	Match Quality	Hierarchy Response	No Hierarchy Response
Intense rainfall	Nature-1	Good match	It's good to think about the influence of rainfall. *Is the fact that Taiwan is located in the convergent plate-boundary region relevant to the occurrence of a debris flow?*	It's good to think about the influence of rainfall. *What issues might be related to shortcomings in public education?*
Improper development of mountains	Social-1	Bad match, creative phrasing	You seem to have some thoughts about social policy and public administration. *What issues might be related to shortcomings in public education?*	It's a good idea to think about policy making. *What king of weather conditions might increase the possibility of a debris flow hazard?*
The current policies do not regulate urban development well	Social-1	Good match	It's a good idea to think about policy making. *What issues might be related to shortcomings in public education?*	It's a good idea to think about policy making. *Is the fact that Taiwan is located in the convergent plate-boundary zone relevant to the occurrence of a debris flow?*
Batman did not come to help	Social-1	Bad match, non-sense student answer	You seem to have some thoughts about social policy and public administration. *Is the public not aware enough about the severity of debris flow hazards?*	It's a good idea to think about policy making. *Is it possible that the location has a particular geological structure that makes it prone to debris flow?*

randomly. We consider this a reasonable comparison condition because without an idea hierarchy, it is not possible to use a strategy such as described above to compensate for poor matches or to select a next tutorial focus that is close to the current focus, since there is no organization of the inventory of idea nodes.

The extended example in Table 1 illustrates how the `is-a` hierarchy is beneficial for the FSM-based feedback selection over a multi-turn interaction. While this is a constructed example, the evaluation presented later in the paper demonstrates that the hierarchy can indeed be used strategically in practice to yield higher quality feedback responses to real student inputs. In the leftmost column of Table 1 we see a sequence of student contributions. To the right of that we see the selected best matching node from the hierarchy as well as an indication of match quality. The two rightmost columns display the resulting comment and tutorial texts for the Hierarchy approach and the No Hierarchy approach respectively. Note that the Comment texts are presented in plain text, and the Tutorial texts are presented in italics.

In the first row, we see a typical student response. Since it matches well to an idea node, and since it is the first contribution in an extended interaction, both approaches generate an acceptable response. Next, however, the student contributes what is an acceptable idea, but with creative phrasing that does not match well to any of the nodes in the hierarchy. Because of this, a poor selection is made for the best matching node. As a result, in the No Hierarchy case, a slightly incoherent comment is produced. In contrast, next to that in the table we see that with the Hierarchy approach, a more abstract comment text that sounds less incoherent in the context is selected. Next, the student contributes a reasonable idea that again matches well to a node in the hierarchy, but it happens to be the same node that matched previously. Thus, we see in the last column of Table 1 that the No Hierarchy approach generates an identical comment to that used previously. While it is appropriate now and follows coherently from what the student has most recently uttered, it sounds awkward because it is identical to what was produced in the previous exchange. In contrast, if we now look in column 4 of Table 1 for the corresponding Hierarchy based feedback, a different comment is produced since this time the match is good, and thus the algorithm does not select a more abstract feedback message. The selected tutorial text is the same since the student has still not responded to this suggestion from the previous turn. In this case, a repeated tutorial text sounds less awkward than a repeated comment text because it is reminding the student of what the student still has not done. In the final student contribution, a nonsense student answer is given. Again, the same best matching node is selected, but this time the match is bad again. The No Hierarchy approach selects the same comment text for the third time. In contrast, the Hierarchy approach reverts back to the more abstract comment and

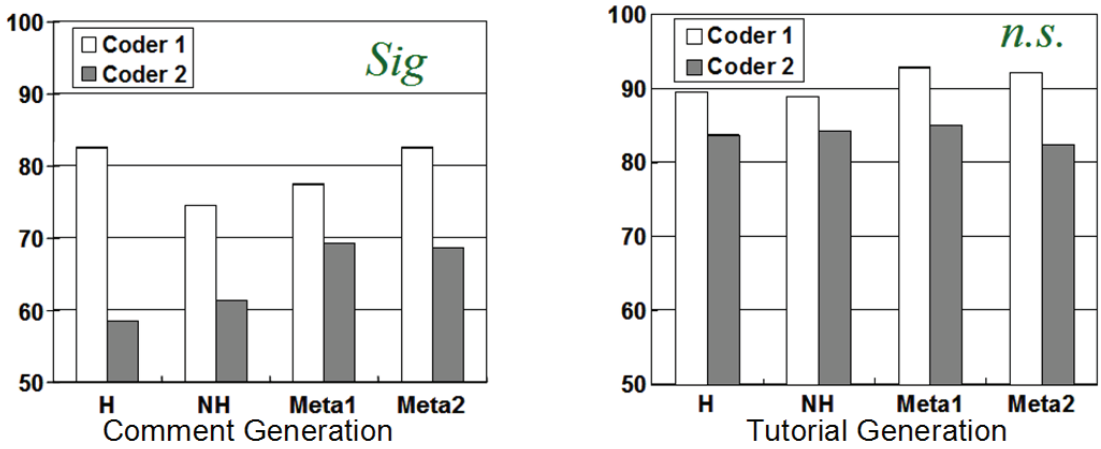

Figure 1 Average ratings for the two human coders for each of the four approaches for Comment selection and Tutorial selection separately.

selects a tutorial with the continued focus on social policy, which has been a coherent theme across the last three exchanges. In contrast, the focus of tutorial texts in the No Hierarchy approach does not follow a logical progression.

4 Evaluation of Baseline Approaches

We first evaluated the two baseline approaches, namely the Hierarchy (H) approach and the No Hierarchy (NH) approach using a corpus containing 163 entries of ideas contributed from 25 Taiwanese high school students in response to the question "What are the possible factors that may cause a debris flow hazard to happen?", which was given to them as part of an inquiry learning unit in their earth sciences class. We refer to this as the debris flow hazard question. For each input entry, the two comment/tutorial selection methods, *Hierarchy* (H) and *No Hierarchy* (NH), were executed in order to generate two feedback versions, each consisting of a Comment and a Tutorial. We recruited two independent judges not directly involved in the research to rate the quality of the comment and tutorial offered by the two different methods in response to each student idea. Each coder assigned a separate binary score (Acceptable/Not_Acceptable) to each comment and tutorial message produced by each of the two methods. In order to prevent biasing the judges in favor of one approach or the other, the coders were kept blind to the method used in constructing each message by presenting the feedback produced by the two approaches as a pair, but in a randomized order. Note that because of this randomization it was necessary for each student contribution/feedback pair to be treated as an independent exchange, divorced from its extended context, for the purpose of this evaluation, although this disadvantages the evaluation of the context-oriented Hierarchy approach.

Figure 1 displays the results not only for the two baseline approaches, but also for the heuristic approaches described below. First, let us consider only the comparison between the two baseline approaches. Coder 1's scores were significantly higher than those of Coder 2, as evaluated using a binary logistic regression ($p < .001$). Nevertheless, we also see a trend for Coder 1 to prefer the Hierarchy method, while with Coder 2 we see the opposite trend. However, neither of these differences were statistically significant. Thus, based on this initial evaluation, the simple version of the hierarchy approach did not yield a significant improvement in feedback generation over the no hierarchy approach. A reliable improvement over the baseline approach is, however, achieved using the optimizations discussed in the remainder of the paper, and evaluated in Section 6.

5 Learning Heuristics for Optimizing Feedback Selection

An error analysis from our initial evaluation revealed that the strategy for selecting a more abstract comment to compensate for a bad match was only effective about half the time. In fact, in some cases, the No Hierarchy response was preferred. Furthermore, we noticed that using the magnitude of semantic similarity as the only basis for determining whether it was better to select a feedback message at a higher level of abstraction was not an effective strategy since the meaning of good versus bad match in terms of semantic similarity score was not stable across idea nodes due to differing amounts of variance in possible phrasings of the concepts and variable coverage of alternative phrasings in the prototype lists.

To capture the regular patterns found in the acceptability evaluation of the human raters, we used a decision tree learning approach. Note that this is similar in spirit to prior work on dialogue strategy induction based on usability questionnaires [Litman et al., 2000]. However, decision tree learning is a simpler machine learning approach to the reinforcement learning approach used in Litman's work, and thus requires far less training data. We achieved encouraging

results with a minimal amount of training data, specifically 163 examples. The amount of data required for our investigation was an amount that can be collected in a single classroom session and evaluated by human raters for the acceptability of the generated feedback within 2 working days. The distribution of examples was consistent with the natural variation in frequency between the occurrence of the alternative ideas represented in the hierarchy. No attempt was made to keep the distribution of ideas equal.

The features we selected for our machine learning approach were extracted from the logs recorded by our feedback generation algorithm discussed above in Section 3. These features are based only on information that one would expect to have access to in a wide range of possible approaches relying on a domain ontology to maintain coherence in the dialogue. Based on our error analysis of the two baseline approaches, we selected as features the idea_id of the closest matching node in the hierarchy, the similarity score based on lexical overlap between the student contribution and the best matching prototype explanation in the best matching idea node, the feedback_message_id of the selected message from the Hierarchy approach, the feedback_message_id of the selected message from the No Hierarchy approach, and a binary variable indicating whether the similarity score was above average or below average.

Because the human coders were not always consistent in their evaluation of the alternative feedback approaches, we trained the decision tree learner to predict whether the average acceptability rating of the two judges would be high or low. We did this by averaging the acceptability ratings of the two raters and then assigning instances with an above median average score to a target score of High, and all others to a target score of Low. We used the J48 implementation of decision tree learning found in the Weka toolkit [Witten and Frank, 2005]. In order to compensate for the high branch factor in several of the features, we set the binarySplits parameter to true. Furthermore, we increased the minimum number of instances per leaf node to 3 in order to avoid over-fitting.

Overall, it was more difficult to predict the acceptability of the tutorial feedback than the comment feedback, partly because the frequency of an unacceptable tutorial selection was relatively rare for both feedback generation approaches. Furthermore, the acceptability of the hierarchy approach was less predictable than that of the no hierarchy approach, partly because the strategy for selecting a more abstract feedback message was sometimes counter-productive. Using cross validation, we were able to achieve a high average performance for Comment selection, and a somewhat lower performance for Tutorial selection for both the Hierarchy and No Hierarchy approaches. For Comment selection, we achieved a percent accuracy of 84% (.67 Kappa) for the No Hierarchy approach and 80% (.61 Kappa) for the Hierarchy approach. For Tutorial selection, we achieved a percent accuracy of 83% (.51 Kappa) for the No Hierarchy approach and 78% (.21 Kappa) for the Hierarchy approach. Despite the relatively low level of predictability of the acceptability judgments, this noisy predictor of acceptability lead to an overall increase in acceptability rating of generated feedback in their use within the meta-heuristics described and evaluated in the next section.

6 Evaluating Heuristic Approaches to Feedback Selection

Because the heuristic methods work by choosing to select either the feedback produced with the Hierarchy or No Hierarchy approach depending upon their prediction of which will optimize their score, we can estimate the human rater's assessment of their respective feedback quality by selecting for each example the corresponding human rating, either from the Hierarchy or No Hierarchy feedback, for each idea depending on which would be selected by the heuristic.

We evaluated the average score for the two meta heuristics using a form of stratified cross-validation as follows: On each iteration, using 90% of the data we trained a decision tree learner separately for the Hierarchy and No Hierarchy approaches just as in the evaluation described in the preceding section to predict whether the quality of the feedback produced with that approach will be rated as *high* or *low*. We used those trained classifiers to assign predictions to the feedback produced by the Hierarchy and No Hierarchy approaches in the 10% of the data set aside as testing data on that iteration. We then evaluated the quality of the feedback produced by the heuristic approaches in the testing data by applying the two meta heuristics to the predicted quality scores. Thus, for Meta 1, if the quality prediction for the Hierarchy approach was low, we would select the No Hierarchy feedback, otherwise we would select the hierarchy feedback. For Meta 2, we would select the No Hierarchy feedback only in the case where the Hierarchy feedback was predicted to be of low quality and the No Hierarchy feedback was predicted to be of high quality. Because the quality of the Hierarchy and No Hierarchy approaches was already rated by the human coders, we could compute the quality of the heuristic approaches based on the already assigned scores.

Figure 1 shows the success of the two heuristic approaches at increasing the overall quality of the generated feedback. Overall, both for Comment and Tutorial generation, the meta-heuristics were more successful than the baseline approaches. We evaluated the statistical significance of this difference separately for Comment and Tutorial generation in each case using a binary logistic regression where the outcome variable was the binary acceptability rating and the two independent variables were the feedback generation approach and the rater. In this statistical model, the acceptability rating of each rater for each feedback approach on each example student contribution was treated as a separate data point. Specifically for Comment selection, the proportion of acceptable to non-acceptable feedback messages was significantly higher for both Meta heuristics in comparison to

the No Hierarchy approach ($p < .05$). The best scoring approach was Meta-heuristic 2, which takes the predicted quality of both the Hierarchy and No Hierarchy approaches into account. For Tutorial selection, the meta heuristics yielded an improvement although the improvement failed to reach the level of statistical significance. However, it should be noted that all approaches performed well from a practical standpoint. Thus, the improvement was mainly necessary for Comment selection, where we see the greatest impact.

7 Conclusions and Current Directions

We have described a new hybrid language processing approach that uses a concept hierarchy to maintain coherence in its feedback selection in the context of an extended interaction with users over multiple turns. In contrast to the Graesser et al. [2001] approach, which is superficially similar to ours, our hierarchy based feedback approach is motivated by general principles of dialogue coherence rather than a specific scaffolded knowledge elicitation strategy consisting of sequences of more and more pointed hints. Specifically, we evaluate a decision tree learning approach for tuning the strategy for using the hierarchy in feedback selection, demonstrating a significant advantage for the optimized approach in comparison with an approach that does not use the hierarchy. These successful results were achieved with a very small amount of training data, in contrast to other work using reinforcement learning approaches that require at least an order of magnitude more data [e.g., Tetreault & Litman, 2006]. These results achieve a practical level of acceptability in feedback quality despite taking a far simpler, less knowledge engineering oriented approach than that of Popescu and colleagues [Popescu, Aleven, & Koedinger, 2003].

In a recent lab study we demonstrated that our feedback generation approach is successful for increasing student learning as students engage in the Debris Flow Hazard task, whether they do so individually or in pairs [Wang et al., submitted].

References

[Aleven et al., 2006] Aleven, V., Ashley, K., Lynch, C., Pinkwart, N. [2006]. Proceedings of the ITS 2006 Workshop on Intelligent Tutoring Systems for Ill-Structured Domains, Springer-Verlag.

[Donmez et al., 2005] Donmez, P., Rose, C. P., Stegmann, K., Weinberger, A., and Fischer, F. [2005]. Supporting CSCL with Automatic Corpus Analysis Technology, to appear in *the Proceedings of Computer Supported Collaborative Learning*.

[Graesser et al., 2001] Graesser, A.C., Person, N., Harter, D., & TRG [2001]. Teaching tactics and dialog in AutoTutor. International Journal of Artificial Intelligence in Education

[Ko & Seo, 2006] Ko, Y., & Seo, J. [2004]. Learning with Unlabeled Data for Text Categorization Using a Bootstrapping and a Feature Projection Technique. *Proceedings of the Association for Computational Linguistics*, pp 255-262

[Litman et al., 2000] Litman, D., Kearns, M., Singh, S. and Walker, M. [2000]. Automatic Optimization of Dialogue Management. In *Proceedings of the 18th International Conference on Computational Linguistics [COLING-2000]*.

[Litman & Pan, 2002] Litman, D. and Pan, S. [2002]. Designing and Evaluating an Adaptive Spoken Dialogue System. *User Modeling and User-Adapted Interaction.* Vol. 12, No. 2/3, pp. 111-137, 2002.

[Ma & Chen, 2003] Ma, W. & Chen, K. [2003]. Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff. *Proceedings of the second SIGHAN workshop on Chinese language processing,* volume 17.

[Malatesta et al., 2002] Malatesta, K., Wiemer-Hastings, P. and Robertson, J. [2002]. Beyond the short answer question with research methods tutor. In *Proceedings of the Intelligent Tutoring Systems Conference.*

[Popescu, Aleven, & Koedinger, 2003] Popescu, O., Aleven, V. and Koedinger, K. [2003]. A kowledge based approach to understanding students explanations. In *Proceedings of the AI in Education Workshop on Tutorial Dialogue Systems: With a View Towards the Classroom,* IOS Press.

[Rosé et al., 2001] Rosé, C. P., Jordan, P., Ringenberg, M., Siler, S., VanLehn, K., Weinstein, A. [2001]. Interactive Conceptual Tutoring in Atlas-Andes, *Proceedings of AI in Education* 2001

[Rosé & VanLehn, 2005] Rosé C. P., & VanLehn, K. [2005]. An Evaluation of a Hybrid Language Understanding Approach for Robust Selection of Tutoring Goals, *International Journal of AI in Education 15[4].*

[Tetrealt & Litman, 2006] Tetreault, J. & Litman, D. [2006]. Using Reinforcement Learning to Build a Better Model of Dialogue State, *Proceedings of the 11th Conference of the European Association for Computational Linguistics*.

[Wang et al., 2006] Wang, H., Li, T., Huang, C., Chang, C., Rosé, C. P. [2006]. VIBRANT: A Brainstorming Agent for Computer Supported Creative Problem Solving, *Proceedings of the Intelligent Tutoring Systems Conference*

[Wang et al., submitted] Wang, H., Rosé, C. P., Chang, C., Huang, C., Cui, Y., & Li, T. [submitted]. Thinking Hard Together: the Long and Short of Collaborative Idea Generation, submitted to *ACM SIG-CHI '07*.

[Witten & Frank, 2005] Witten, I. H. & Frank, E. [2005]. *Data Mining: Practical Machine Learning Tools and Techniques*, second edition, Elsevier: San Francisco