Predicting Learning in Tutoring with the Landscape Model of Memory

Arthur Ward
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, Pa., 15260, USA
artward@cs.pitt.edu

Diane Litman
Learning Research and Development Center
University of Pittsburgh
Pittsburgh, Pa., 15260, USA
litman@cs.pitt.edu

Abstract

A Landscape Model analysis, adopted from the text processing literature, was run on transcripts of tutoring sessions, and a technique developed to count the occurrence of key physics points in the resulting connection matrices. This point-count measure was found to be well correlated with learning.

1 Introduction

Human one-to-one tutoring often yields significantly higher learning gains than classroom instruction (Bloom, 1984). This difference motivates natural language tutoring research, which hopes to discover which aspects of tutorial dialogs correlate with learning. Much of this research focuses on various dialog characteristics. For example, (Graesser et al., 1995) argue that the important components of tutoring include question answering and explanatory reasoning. In other work (Litman et al., 2004) examine dialog characteristics that can be identified automatically, such as ratio of student to tutor words, and average turn length.

In this paper, rather than look at characteristics of the tutoring dialog itself, we feed the dialog into a computational model of student memory, in which we then find a measure correlated with learning. This "Landscape Model" (van den Broek et al., 1996) proves useful for predicting how much students remember from tutoring sessions, as measured by their learning gains.

We will first briefly describe the Landscape Model. Then we will describe the tutoring experiments from which we draw a corpus of dialogs, and how the model was applied to this corpus. Finally, we cover the model's success in predicting learning.

2 The Landscape Model

The Landscape Model was designed by van den Broek et al. (1996) to simulate human reading comprehension. In this model, readers process a text sentence-by-sentence. Each sentence contains explicitly mentioned concepts which are added into working memory. In addition, the reader may re-instantiate concepts from earlier reading cycles or from world knowledge in an effort to maintain a coherent representation. Concepts are entered into working memory with initial activation values, which then decay over subsequent reading cycles.

After concepts are entered, the model calculates connection strengths between them. Two concepts that are active in working memory at the same time will be given a link. The higher the levels of concept activation, the stronger the link will be. Van den Broek et al. (1996) give this formula for calculating link strengths: $\sum_{i=1}^{l} A_{xi} A_{yi}$

This defines the strength of the connection between concepts x and y as the product of their activations (A) at each cycle i, summed over all reading cycles.

Two matrices result from these calculations. The first is a matrix of activation strengths, showing all the active concepts and their values for each reading cycle. The second is a square matrix of link values showing the strength of the connection between

Proceedings of the 2nd Workshop on Building Educational Applications Using NLP,
pages 21–24, Ann Arbor, June 2005. ©Association for Computational Linguistics, 2005

each pair of concepts. Van den Broek et al. (1996) demonstrate a method for extracting a list of individual concepts from these matrices in order of their link strengths, starting with the strongest concept. They show a correlation between this sequence and the order in which subjects name concepts in a free-recall task.

In van den Broek's original implementation, this model was run on short stories. In the current work, the model is extended to cover a corpus of transcripts of physics tutoring dialogs. In the next section we describe this corpus.

3 Corpus of Tutoring Transcripts

Our corpus was taken from transcripts collected for the ITSPOKE intelligent tutoring system project (Litman and Silliman, 2004). This project has collected tutoring dialogs with both human and computer tutors. In this paper, we describe results using the human tutor corpus.

Students being tutored are first given a pre-test to gauge their physics knowledge. After reading instructional materials about physics, they are given a qualitative physics problem and asked to write an essay describing its solution. The tutor (in our case, a human tutor), examines this essay, identifies points of the argument that are missing or wrong, and engages the student in a dialog to remediate those flaws. When the tutor is satisfied that the student has produced the correct argument, the student is allowed to read an "ideal" essay which demonstrates the correct physics argument. After all problems have been completed, the student is given a post-test to measure overall learning gains. Fourteen students did up to ten problems each. The final data set contained 101,181 student and tutor turns, taken from 128 dialogs.

4 Landscape Model & Tutoring Corpus

Next we generated a list of the physics concepts necessary to represent the main ideas in the target solutions. Relevant concepts were chosen by examining the "ideal" essays, representing the complete argument for each problem. One hundred and twelve such concepts were identified among the 10 physics problems. Simple keyword matching was used to identify these concepts as they appeared in each line of the dialog. A small sample of these concepts and their keywords is shown in Table 1.

Concept Name	Keywords
above	above, over
acceleration	acceleration,accelerating
action	action, reaction
affect	experience,experienced
after	after, subsequent
air friction	air resistance, wind resistance
average	mean
ball	balls, sphere
before	before, previous
beside	beside, next to

Table 1: Examples of concepts and keywords

Each concept found was entered into the working memory model with an initial activation level, which was made to decay on subsequent turns using a formula modeled on van den Broek (1996). Concept strengths are assumed to decay by 50% every turn for three turns, after which they go to zero. A sample portion of a transcript showing concepts being identified, entering and decaying is shown in Table 2. Connections between concepts were then calculated as described in section two. A portion of a resulting concept link matrix is shown in Table 3.

It should be noted that the Landscape model has some disadvantages in common with other bag-of-words methods. For example, it loses information about word order, and does not handle negation well.

As mentioned in section two, van den Broek et al. created a measure that predicted the order in which individual concepts would be recalled. For our task, however, such a measure is less appropriate. We are less interested, for example, in the specific order in which a student remembers the concepts "car" and "heavier," than we are in whether the student remembers the whole idea that a heavier car accelerates less. To measure these constellations of concepts, we created a new measure of idea strength.

5 Measuring Idea Strength

The connection strength matrices described above encode data about which concepts are present in each dialog, and how they are connected. To extract useful information from these matrices, we used the idea of a "point." Working from the ideal essays, we identified a set of key points important for the solution of each physics problem. These key points

Turn	Text	Concepts			
		car	heavier	acceleration	cause
Student	I don't know how to answer this it's got to be slower, cause, it's the car is heavier but	5	5	0	0
Tutor	yeah, just write whatever you think is appropriate	2.5	2.5	0	0
Student	ok,	1.25	1.25	0	0
Essay	The rate of acceleration will decrease if the first car is towing a second, because even though the force of the car's engine is the same, the weight of the car is double	5	0.625	5	5
Student	ok	2.5	0	2.5	2.5
Tutor	qualitatively,um, what you say is right, you have correctly recognized that the force, uh, exerted will be the same in both cases,uh, now, uh, how is force related to acceleration?	1.25	0	5	1.25

Table 2: Portion of a transcript, showing activation strengths per turn

	car	heavier	acceleration	cause	decelerates	decrease
car	0	35.9375	115.234375	102.34375	33.203125	33.2
heavier	0	0	3.125	3.125	3.125	3.13
acceleration	0	0	0	107.8125	42.1875	42.19
cause	0	0	0	0	33.203125	33.2
decelerates	0	0	0	0	0	66.41
decrease	0	0	0	0	0	0

Table 3: Portion of link value table, showing connection strengths between concepts

are modeled after the points the tutor looks for in the student's essay and dialog. For example, in the "accelerating car" problem, one key point might be that the car's acceleration would decrease as the car got heavier. The component concepts of this point would be "car," "acceleration," "decrease," and "heavier." If this point were expressed in the dialog or essay, we would expect these concepts to have higher-than-average connection strengths between them. If this point were not expressed, or only partially expressed, we would expect lower connection strengths among its constituent concepts.

The strength of a point, then, was defined as the sum of strengths of all the links between its component concepts. Call the point in the example above "p_i." point p_i has n = 4 constituent concepts, and to find its strength we would sum the link strengths between their pairs: "car-acceleration," "car-decrease," "car-heavier," "acceleration-decrease,", "acceleration-heavier," and "decrease-heavier." Using values from Table 3, the total strength for the point would therefore be:

$$pointStr_{p_in} = 115.23 + 33.2 + 35.94 + 42.19 + 3.13 + 3.13 = 232.81.$$

For each point, we determined if its connections were significantly stronger than the average. We generate a reference average $AvgStr_n$ by taking 500 random sets of n concepts from the same dialog and averaging their link weights, where n is the number of concepts in the target point [1]. If the target point was found to have a significantly ($p < .05$ in a t-test) larger value than the mean of this random sample, that point was above threshold, and considered to be present in the dialog.

The number of above-threshold points was added up over all dialogs for each student. The total point-count for student S is therefore:

$$pointCount_S = \sum_{i=1}^{P} T(pointStr_{p_in}, AvgStr_n)$$

Where P is the total number of points in all dialogs, and T is a threshold function which returns 1 if $pointStr_{p_in} > AvgStr_n$, and 0 otherwise.

Fifty-seven key points were identified among the ten problems, with each point containing between two and five concepts. The next section describes how well this point-count relates to learning.

[1] 500 was chosen as the largest feasible sample size given runtime limitations

6 Results: Point Counts & Learning

We first define "concept-count" to be the number of times physics concepts were added to the activation strength matrix. This corresponds to each "5" in Table 2. Now we look at a linear model with post-test score as the dependant variable, and pre-test score and concept-count as independent variables. In this model pre-test score is significant, with a p-value of .029, but concept-count is not, with a p-value of .270. The adjusted R squared for the model is .396

Similarly, in a linear model with pre-test score and point-count as independent variables, pre-test score is significant with a p-value of .010 and point-count is not, having a p-value of .300. The adjusted R squared for this model is .387.

However, the situation changes in a linear model with pre-test score, concept-count and point-count as independent variables, and post-test score as the dependent variable. Pre-test is again significant with a p-value of .002. Concept-count and point-count are now both significant with p-values of .016 and .017, respectively. The adjusted R-squared for this model rises to .631.

These results indicate that our measure of points, as highly associated constellations of concepts, adds predictive power over simply counting the occurrence of concepts alone. The number of concept mentions does not predict learning, but the extent to which these concepts are linked into relevant points in the Landscape memory model is correlated with learning.

7 Discussion

Several features of the resulting model are worth mentioning. First, the Landscape Model is a model of memory, and our measurements can be interpreted as a measure of what the student is remembering from the tutoring session taken as a whole.

Second, the point-counts are taken from the entire dialog, rather than from either the tutor or student's contributions. Other results suggest that it would be interesting to investigate the extent to which these points are produced by the student, the tutor, or both...and what effect their origin might have on their correlation with learning. For example, (Chi et al., 2001) investigated student-centered, tutor-centered and interactive hypotheses of tutoring and found that students learned just as effectively when tutor feedback was suppressed. They suggest, among other things, that students self-construction of knowledge was encouraging deep learning.

8 Summary and Future Work

We have shown that the Landscape Model yields a measure significantly correlated with learning in our human-human tutoring corpus. We hope to continue this work by investigating the use of well researched NLP methods in creating the input matrix. In addition, machine learning methods could be used to optimize the various parameters in the model, such as the decay rate, initial activation value, and point strength threshold.

9 Acknowledgments

We thank Tessa Warren, Chuck Perfetti and Franz Schmalhofer for early advice on this project. This research is supported by ONR (N00014-04-1-0108).

References

B. Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13:4–16.

M. Chi, S. Siler, H. Jeong, T. Yamauchi, and R. Hausman. 2001. Learning from human tutoring. *Cognitive Science*, 25:471–533.

A. Graesser, N. Person, and J. Magliano. 1995. Collaborative dialog patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology*, 9:359–387.

D. Litman and S. Silliman. 2004. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Companion Proc. of the Human Language Technology Conf: 4th Meeting of the North American Chap. of the Assoc. for Computational Linguistics*.

Diane J. Litman, Carolyn P. Rosé, Kate Forbes-Riley, Kurt VanLehn, Dumisizwe Bhembe, and Scott Silliman. 2004. Spoken versus typed human and computer dialogue tutoring. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems(ITS)*. Maceio, Brazil.

P. van den Broek, K. Risden, C.R. Fletcher, and R. Thurlow. 1996. A landscape view of reading: Fluctuating patterns of activation and the construction of a stable memory representation. In B.K. Britton and A.C. Graesser, editors, *Models of understanding text*, pages 165–187. Mahweh, NJ: Lawrence Erlbaum Associates.