Predicting Change in Student Motivation by Measuring Cohesion between Tutor and Student

Arthur Ward
Department of Biomedical Informatics
University of Pittsburgh
Pittsburgh, Pa., 15232
akw13@pitt.edu

Diane Litman
Department of Computer Science, LRDC
University of Pittsburgh
Pittsburgh, Pa., 15260
litman@cs.pitt.edu

Maxine Eskenazi
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, Pa., 15213
max@cmu.edu

Abstract

We apply a previously reported measure of dialog cohesion to a corpus of spoken tutoring dialogs in which motivation was measured. We find that cohesion significantly predicts changes in student motivation, as measured with a modified MSLQ instrument. This suggests that non-intrusive dialog measures can be used to measure motivation during tutoring.

1 Introduction

Motivation is widely believed to be an important factor in learning, and many studies have found relationships between motivation and educational outcomes. For example Pintrich and DeGroot (1990) found that students' motivational state was a significant predictor of classroom performance. In addition, pedagogically significant behaviors such as dictionary lookup in the REAP (Brown and Eskenazi, 2004) vocabulary tutor have been shown to be positively correlated with motivation assessments (DelaRosa and Eskenazi, 2011). Also, in a separate study with the REAP tutor, attempts to manipulate reading motivation by presenting more interesting stories were shown to improve vocabulary learning (Heilman et al., 2010).

In addition to influencing learning outcomes, motivational state may also affect which interventions will be effective during tutoring. For example, Ward and Litman (2011) have shown that motivation can significantly affect which students benefit from a reflective reading following interactive tutoring with a the Itspoke (Litman and Silliman, 2004) tutor.

An accurate way to measure student motivation during tutoring could therefore be valuable to Intelligent Tutoring System (ITS) researchers. Several self-report instruments have been developed which measure various aspects of motivation (e.g. (Pintrich and DeGroot, 1990; McKenna and Kear, 1990)). However, these instruments are too intrusive to be administered *during* tutoring, for fear of fatally disrupting learning. We would prefer a non-intrusive measure which would allow an ITS to detect when student motivation is decreasing so as to launch a motivational intervention. Similarly, the ITS should be able to detect when motivation is increasing again, to determine if the intervention worked. As mentioned above, such a measure might also allow an ITS to determine when it would be effective to use certain instructional tactics.

In this work we investigate *cohesion* as a non-intrusive measure of motivation for natural language dialog based ITS. As defined more precisely below, our measure of cohesion quantifies lexical and semantic similarity between tutor and student dialog utterances. We hypothesize that this measure of lexical similarity may be related to motivation in part because other measures of dialog similarity have been shown to be related to task success. For example, there is evidence that perceived similarity between a student's own speech rate and that of a recorded task request increases the student's feelings of immediacy, which are in turn linked to greater compliance with the request to perform a task (Buller and Aune, 1992). [1] In addition, Ward and Litman (2006; 2008) investigated a measure of lexical similarity between

[1] In this experiment, the task was to watch a series of videos.

Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, pages 136–141,
Portland, Oregon, 24 June 2011.

the tutor and student partners in a tutoring dialog which was shown to be correlated with task success in several corpora of tutorial dialogs.

Measures of cohesion have also been used in a variety of NLP tasks such as measuring text readability (e.g. (Pitler and Nenkova, 2008)), measuring stylistic differences in text (Mccarthy et al., 2006), and for topic segmentation in tutorial dialog (Olney and Cai, 2005).

Given the previously mentioned results relating motivation to educational task success, these links between task success and cohesion lead us to hypothesize a direct correlation between motivation and cohesion when using the Itspoke tutor.

We will first briefly describe the Itspoke tutor, and the corpus of tutoring dialogs used in this study. We will then describe the instrument we used to measure motivation both before and immediately after tutoring, then we will describe the algorithm used to measure cohesion in the tutoring dialogs. Finally, we show results of correlations between the measure of motivation and the measure of cohesion. We will find that the *change* in motivation is significantly correlated with dialog cohesion.

2 Itspoke System and Corpus

Itspoke (**I**ntelligent **T**utoring **SPOKE**n dialog system) is a spoken dialog tutoring system which teaches qualitative physics. It provides a spoken dialog interface to the Why2-Atlas (VanLehn et al., 2002) tutor, and has recently been re-implemented using the TuTalk (Jordan et al., 2007) dialog platform. The Itspoke tutor presents a problem in qualitative physics, and asks the student an initial question. The student answers the question, and the dialog continues until all points have been covered to the tutor's satisfaction.

The corpus used in the current work was collected in a previous study (Ward and Litman, 2011), using novice subjects who had never taken a college physics course. Before tutoring, students were given a motivation survey which will be described in Section 3. They then engaged Itspoke in five tutoring dialogs as described above. Immediately after tutoring they were given the motivation questionnaire again, with tenses changed as appropriate.

Speaker	Utterance
Tutor	To see which vehicle's **change** in motion is greater, we use the definition of **acceleration**. What is the definition of acceleration?
Student	**Change** in **velocity**.
Tutor	Excellent. Acceleration is defined as the amount **velocity changes** per unit time.

Table 1: Dialog turns, with Token, Stem, and Semantic Similarity Matches in **bold** (as discussed in Section 4).

166 subjects were recruited by flyer, by advertisement during an undergraduate psychology course, or from the University of Pittsburgh's psychology subject pool. Of these, 40 were dismissed after pretest as "middle third" knowledge students, following extreme groups design (Feldt, 1961). Extreme groups design was adopted to increase the power of a "high" vs "low" knowledge comparison, which is reported elsewhere (Ward, 2010). Another 27 students were not used for various reasons including incomplete data. This left a corpus of 99 subjects who each participated in 5 tutorial dialogs.

Table 1 shows an exchange from one of these dialogs. The tutor asks a question about the current problem, which the student then answers. The tutor restates the answer, and (later in the dialog) proceeds on to the next point of discussion.

3 Motivation Measure

In this study we measure motivation using a reduced version of the "Motivated Strategies for Learning Questionnaire (MSLQ)" developed by Pintrich and DeGroot (1990). This version of the MSLQ is also based on work by Roll (2009), who adapted it for use in an IPL (Invention as Preparation for Learning (Schwartz and Martin, 2004)) tutoring environment. Our motivational survey is shown in Figure 1.

Questions one and two address "self-regulation," particularly the students' tendency to manage and control their own effort. Question one is on a reversed scale relative to the other questions, so responses to it were inverted. Question three addresses "self-efficacy," the students' expectation of success on the task. Questions four and five address "intrinsic value," the students' beliefs about the importance and interest of the task.

These dimensions of motivation are theoretically

Please read the following statements and then click a number on the scale that best matches how true it is of you. 1 means “not at all true of me” whereas 7 means “very true of me”.

1. I think that when the tutor is talking I will be thinking of other things and won’t really listen to what is being said.
2. If I could take as much time as I want, I would spend a lot of time on physics tutoring sessions.
3. I think I am going to find the physics tutor activities difficult.
4. I think I will be able to use what I learn in the physics tutor sessions in my other classes.
5. I think that what I will learn in the physics tutor sessions is useful for me to know.

Figure 1: Pre-tutoring Motivational Survey

distinct. However, except for question three (the self-efficacy question), responses to these questions were all very significantly correlated with each other in our survey ($p < .01$).

Table 2 shows values of Cronbach’s Alpha (Cronbach, 1951) for various subsets of the motivation questions. Alpha measures the internal consistency of responses to a multi-point questionnaire, and is a function of the number of test items and the correlation between them. Higher values are thought to indicate that the various test items are measuring the same underlying latent construct. For this study we omit Question 3, maximizing Alpha at .716. This is just above the commonly accepted (Gliem and Gliem, 2003) threshold for reliability in such an instrument.

As mentioned above, this instrument was administered both before and (with suitable tense changes) immediately after tutoring. We will report correlations using mean scores on the pre- and post-tutoring measures, as well as for the *change*-in-motivation score, calculated as post-minus-pre.

Questions	**Alpha**
1, 2, 3, 4, 5	0.531
1, 2, 4, 5	0.716
2, 4, 5	0.703
4, 5	0.683

Table 2: Alpha reliability scores for various subsets of questions.

4 Semantic Cohesion Measure

In this work we measure cohesion between tutor and student using the “semantic cohesion” measure first reported by Ward and Litman (2008). This measure counts the number of “cohesive ties” (Halliday and Hasan, 1976) between adjacent tutor and student dialog turns. A cohesive tie can be the repetition of an exact word or word stem, or the use of two words with similar meanings in adjacent turns. Stop words are excluded, and cohesive ties are counted in both the student-to-tutor and the tutor-to-student directions. For example, in the dialog shown in Table 1, the final tutor turn repeats the word “velocity” from the previous student turn. This repetition would be counted as an *exact* cohesive tie. Similarly, the tutor uses the word “changes” following the student’s use of “change.” This would be counted as a stem repetition cohesive tie.

Finally, the student’s use of “velocity” will be counted as a cohesive tie because of its semantic similarity to “acceleration,” from the preceding turn. The algorithm therefore counts four ties in Table 1. As described more completely in (Ward and Litman, 2008), semantic similarity cohesive ties are counted by measuring two words’ proximity in the WordNet (Miller et al., 1990) hierarchy. We use a simple path distance similarity measure, as implemented in NLTK (Loper and Bird, 2002). This measure counts the number of edges N in the shortest path between two words in WordNet, and calculates similarity as 1 / (1 + N). Our implementation of this semantic similarity measure allows setting a threshold θ, such that only word pairs with stronger-than-threshold similarity are counted. Table 3 shows some semantic similarity pairs counted with a threshold of 0.3.

We obtain a normalized cohesion score for each dialog by dividing the tie count by the number of turns in the dialog. We then sum the line normalized counts over all the dialogs for each student, resulting in a per-student cohesion measure.

motion-contact
man-person
decrease-acceleration
acceleration-change
travel-flying

Table 3: Example Semantic ties: $\theta = 0.3$

5 Results

We ran correlations between the change-in-motivation score described in Section 3 and the semantic similarity measure of cohesion described in Section 4. We report results for a semantic similarity threshold of .3 for consistency with (Ward and Litman, 2008), however the pattern of results is not sensitive to this threshold. Significant results were obtained for all thresholds between .2 and .5, in .1 increments. [2] In addition, we report results for the motivation measure with the third question removed for consistency with (Ward and Litman, 2011). However the pattern of results is not sensitive to this exclusion, either. Significant results were also obtained using the entire questionnaire.

Motivation Measure	Cor.	pValue
pre-Tutoring	0.02	0.86
Change	0.21	**0.03**
post-Tutoring	0.19	0.055

Table 4: Cohesion - Motivation Correlations. N = 99. $\theta = 0.3$

In all cases, the change in motivation was found to be significantly and positively correlated with the cohesiveness of the tutoring dialog. More lexical similarity between tutor and student was predictive of increased student motivation. As shown in the middle row of Table 4, the correlation with motivational change, using a threshold of .3 and the reduced motivation measure was r(97)= .21, p = 0.03.

Interestingly, as shown in the top and bottom rows of Table 4, neither motivation before tutoring r(97) = .02, p=.86, nor after tutoring r(97) = .19, p = .055, was significantly correlated with cohesion, although the post-tutoring measure achieves a strong trend.

Pre- and post-tutoring mean motivation levels were, however, significantly correlated with each other (R(97) = .69, $p < .0001$). Mean motivation levels also showed a non-significant improvement from 4.31 before tutoring to 4.44 after tutoring.

6 Discussion and Future Work

We have brought forward evidence that cohesion in tutorial dialog, as measured in this paper, is correlated with changes in student motivation. This suggests that dialog cohesion may be useful as a non-intrusive measure of motivational fluctuations.

As discussed in Section 1, other researchers have investigated various types of cohesion, and their relationship to things such as task success and learning. In addition, work has been done investigating the role of motivation in learning. However, we believe ours is the first work relating dialog cohesion directly to user motivation.

The presence of a correlation between cohesion and motivation leaves open the possibility that more motivated students are experiencing greater task success in the tutor, and so generating more cohesive dialogs. [3] Note, however, that the very non-significant correlation between *pre*-dialog motivation and dialog cohesion argues against this possibility. Instead, it seems that some process is both creating dialog cohesion and *improving* student motivation. The lack of significance in the post-dialog/motivation correlation may be due to data sparsity.

In future work, we hope to investigate other dialog features which may be even better predictors of student motivation. As mentioned in Section 1, we became interested in dialog similarity metrics partly because of their association with task success. These kinds of associations between task success and dialog have also been shown for dialog *entrainment*.

In this discussion we will use the term "entrainment" for the phenomenon in which conversational partners' speech features become more similar to each other at many levels, including word choice, over the course of a dialog. [4] As mentioned above, we use the term "cohesion" for *overall* similarity of word choice between speakers in a dialog, perhaps resulting from entrainment.

Users appear to entrain strongly with dialog systems. For example, Brennan (1996) has found that users are likely to adopt the terms used by a WOZ dialog system, and that this tendency is at least as strong as with human dialog partners. Similarly, Parent and Eskenazi (2010) showed that users of the Let's Go (Raux et al., 2005) spoken dialog system quickly entrain to its lexical choices.

[2] Note from the path distance formula that thresholds between .5 and 1 are impossible

[3] We thank an anonymous reviewer for prompting this discussion.

[4] This definition conflates studies of priming, alignment, convergence and accommodation.

As with measures of dialog similarity, dialog entrainment has been found to be related to satisfaction and success in task oriented dialogs. For example, Reitter and Moore (2007) found that lexical and syntactic repetition predicted task success in the MapTask corpus. Similarly, Ward and Litman (2007) found that lexical and acoustic-prosodic entrainment are correlated with task success in the Itspoke dialog system. Interestingly, in that work entrainment was more strongly correlated with task success than a measure of dialog cohesion similar to the one used in the current paper. This raises the question of whether such a measure of dialog entrainment might also be a better predictor of motivation than the current measure of cohesion. We hope in future work to further investigate this possibility.

Finally, because we are interested in predicting motivation during tutoring, our dialog metrics may be improved by making them sensitive to the educational domain. For example, exploratory work with our tutor has suggested that a measure of cohesion which only counts cohesive ties between physics terms is better correlated with certain measures of learning than a measure which counts non-physics terms. This suggests that measures of cohesion or entrainment which recognize educational domain words may also improve correlations with motivation.

Acknowledgments

This work was supported by the ONR (N00014-07-1-0039), by the NSF (0631930 and 0914615), and by the LRDC and ISP at the University of Pittsburgh.

References

Susan E. Brennan. 1996. Lexical entrainment in spontaneous dialog. In *International Symposium on Spoken Dialog*, pages 41–44.

Jonathan Brown and Maxine Eskenazi. 2004. Retrieval of authentic documents for reader-specific lexical practice. In *In Proceedings of InSTIL/ICALL Symposium*.

David Buller and R.Kelly Aune. 1992. The effects of speech rate similarity on compliance: Application of communication accommodation theory. *Western Journal of Communication*, 56:37–53.

Lee Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Kevin DelaRosa and Maxine Eskenazi. 2011. Self-assessment of motivation: Explicit and implicit indicators of l2 vocabulary learning. *Proceedings 15th International Conference on Artificial Intelligence Education (AIED)*.

Leonard Feldt. 1961. The use of extreme groups to test for the presence of a relationship. *Psychometrika*, 26(3):307–316.

Joesph Gliem and Rosemary Gliem. 2003. Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales. *Midwest Research to Practice in Adult, Continuing and Community Education*.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. English Language Series. Pearson Education Limited.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, Maxine Eskenazi, Alan Juffs, and Lois Wilson. 2010. Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education*, 20:73–98, January.

Pamela Jordan, Brian Hall, Michael Ringenberg, Yui Cui, and Carolyn Rosé. 2007. Tools for authoring a dialogue agent that participates in learning studies. In *Proc. of Artificial Intelligence in Ed., AIED*, pages 43–50.

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*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Philip M. Mccarthy, Gwyneth A. Lewis, David F. Dufty, and Danielle S. Mcnamara. 2006. Analyzing writing styles with coh-metrix. In *In Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS*.

M.C. McKenna and D.J. Kear. 1990. Measuring attitude toward reading: A new tool for teachers. *The Reading Teacher*, 43(8):626–639.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography (special issue)*, 3 (4):235–312.

Andrew Olney and Zhiqiang Cai. 2005. An orthonormal basis for topic segmentation in tutorial dialog. In *Pro-*

ceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 971–978. Vancouver, October.

Gabriel Parent and Maxine Eskenazi. 2010. Lexical entrainment of real users in the let's go spoken dialog system. In *Proceedings Interspeech-2010*, pages 3018 – 3021, Makuhari, Chiba, Japan.

Paul Pintrich and Elisabeth DeGroot. 1990. Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1):33–40.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 186 – 195.

Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. 2005. Lets go public! taking a spoken dialog system to the real world. In *Proceedings Interspeech-2005*, pages 885 – 888, Lisbon, Portugal.

David Reitter and Johanna D. Moore. 2007. Predicting success in dialogue. In *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 808 – 815, Prague, Czech Republic.

Ido Roll. 2009. *Structured Invention Tasks to Prepare Students for Future Learning: Means, Mechanisms, and Cognitive Processes*. Doctor of philosophy, Carnegie Mellon University, 5000 Forbes Ave. Pittsburgh, Pa.

Daniel Schwartz and Taylor Martin. 2004. Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, 22:129 – 184.

K. VanLehn, P. Jordan, C. Rose, D. Bhembe, M. Boettner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, and Srivastava R. 2002. The architecture of why2-atlas: A coach for qualitative physics essay writing. In *Proc. 6th Int. Conf. on Intelligent Tutoring Systems*, pages 158–167.

Arthur Ward and Diane Litman. 2006. Cohesion and learning in a tutorial spoken dialog system. In *Proceedings of the 19th International FLAIRS Conference (FLAIRS-19)*, pages 533–538, May.

Arthur Ward and Diane Litman. 2007. Dialog convergence and learning. In *Proceedings 13th International Conference on Artificial Intelligence Education (AIED)*, Los Angeles, Ca.

Arthur Ward and Diane Litman. 2008. Semantic cohesion and learning. In *Proceedings 9th International Conference on Intelligent Tutoring Systems (ITS)*, pages 459–469, Ann Arbor, June.

Arthur Ward and Diane Litman. 2011. Adding abstractive reflection to a tutorial dialog system. *Proceedings 24th International FLAIRS (Florida Artificial Intelligence Research Society) Conference*.

Arthur Ward. 2010. *Reflection and Learning Robustness in a Natural Language Conceptual Physics Tutoring System*. Doctor of philosophy, University of Pittsburgh, Pittsburgh, PA. 15260.