Towards Assessing Students' Prior Knowledge From Tutorial Dialogues

Dan Stefanescu
University of Memphis
365 Innovation Drive
Memphis, TN, 38152
dstfnscu@memphis.edu

Vasile Rus
University of Memphis
365 Innovation Drive
Memphis, TN, 38152
vrus@memphis.edu

Arthur C. Graesser
University of Memphis
365 Innovation Drive
Memphis, TN, 38152
graesser@memphis.edu

ABSTRACT

This paper describes a study which is part of a project whose goal is to detect students' prior knowledge levels with respect to a target domain based solely on characteristics of the natural language interaction between students and a state-of-the-art conversational Intelligent Tutoring System (ITS). We report results on dialogues collected from two versions of the intelligent tutoring system DeepTutor: a micro-adaptive-only version and a fully-adaptive (micro- and macro-adaptive) version. We extracted a variety of dialogue and session interaction features including time on task, student-generated content features (e.g., vocabulary size or domain specific concept use), and pedagogy-related features (e.g., level of scaffolding measured as number of hints). We present which of these features are best predictors of pre-test scores as measured by multiple-choice questions.

Keywords

Intelligent Tutoring Systems, Knowledge Assessment, Tutoring Dialogues

1. INTRODUCTION

One-on-one tutoring is one of the most effective forms of instruction [14, 23, 26] because it opens up the opportunity of maximizing learning gains by tailoring instruction to each individual learner. A necessary step towards instruction adaptation is assessing students' knowledge such that appropriate instructional tasks (macro-adaptation) are selected and appropriate feedback is provided while students are working on a particular task (micro-adaptation or within-task adaptation). Students' knowledge state is a moving target and therefore, continuous monitoring and updating is necessary which makes the assessment task quite challenging. We focus in this paper on assessing students' knowledge at the moment when they first start interacting with an Intelligent Tutoring System (ITS), which is a special case of the large problem of assessing students' knowledge state, i.e. mental model. Students' prior knowledge with respect to a target domain is typically assessed using multiple choice pre-tests although other forms of assessment may be used.

Assessing students' prior knowledge is very important task in ITSs because it serves two purposes: enabling macro-adaptivity in ITSs [14], and, when paired with a post-test, establishing a baseline from which the student progress is gauged by computing learning gains. While the role of pre-test is important for macro-adaptivity and for measuring learning gains, a major challenge is the fact that the pre-test and the post-test usually take up time from actual learning. Pre-test may even have a tiring effect on students. Last but not least, designing good pre- and post-tests requires domain expertise and could be an expensive effort. Being able to infer students' knowledge directly from their performance would eliminate the pre-test thus saving time for more training, getting rid of the tiring effects or testing anxieties, and reducing developers' effort.

In this paper, we focus on identifying the most important dialogue features that best correlated with students' prior knowledge as measured by pre-tests consisting of multiple-choice questions. This work is part of a large effort which has two major research goals: (1) to understand to what extent we can predict students' pre-training knowledge levels from dialogue features and (2) what is the minimum length of dialogue that is sufficient for predicting students' knowledge states/levels with good accuracy.

An interesting aspect of our work is the fact that we assess students' knowledge levels from training sessions collected from two versions of the intelligent tutoring system DeepTutor: a micro-adaptive-only version and a fully-adaptive version (the fully adaptive is both macro- and micro-adaptive). Micro-adaptivity is about within-task adaptation: the capacity of the system to select appropriate feedback at every single step while the student is working on an instructional task. The fully-adaptive condition adds macro-adaptivity on top of micro-adaptivity; macro-adaptivity is about selecting and sequencing an appropriate set of tasks to each individual student based on her knowledge level. We used a macro-adaptivity method based on an Item-Response Theory approach [15]. As such, in the micro-adaptive-only version of DeepTutor, students worked on tasks following the one-size-fits-all approach, while in the fully-adaptive condition, 4 different task sequences were assigned to students based on their knowledge levels: low, medium-low, medium-high, and high. Analyzing the data from both conditions, our goal is to identify dialogue features and models based on these features that are good predictors of students' prior knowledge as measured by pre-test scores.

2. RELATED WORK

The most directly relevant work to ours is the one by Lintean and colleagues [10] who studied the problem of inferring students' prior knowledge in the context of an ITS that monitored and scaffolded students' meta-cognitive skills. They compared student-articulated prior knowledge activation (PKA) paragraphs to expert-generated paragraphs or to a taxonomy of concepts related to the target domain (i.e. human circulatory system). Students' prior knowledge levels were modeled as a set of 3 categories: low, medium, and high mental models. There are significant differences between the two approaches. First, we deal with dialogues as opposed to explicitly elicited prior knowledge paragraphs. Second, we do not have access to gold standard paragraphs or correct answers or a taxonomy of concepts that would allow us to make direct comparisons. Third, we model students' prior knowledge using their score on a multiple-choice pre-test.

Predicting students' learning and satisfaction is another area of research relevant to our work, and one of the earliest and most useful applications of Educational Data Mining [cf. 13]. Forbes-Riley and Litman [6] used the PARADISE framework [24] to develop models for predicting student learning and satisfaction [7]. They used 3 types of features: system specific, tutoring specific, and user-affect-related. They used the whole training session as a unit of analysis, which is different from our analysis, in which the units are instructional tasks, i.e. Physics problems. Also, their work was in the context of a spoken dialogue system, while ours focuses on a text/chat-based conversational ITS. In addition, they focused on user satisfaction and learning, while we are interested in identifying students' prior knowledge.

Williams and D'Mello [25] worked on predicting the quality of student answers to human tutor questions, based on dictionary-based dialogue features previously shown to be good detectors of cognitive processes [cf. 25]). To extract these features, they used LIWC (Linguistic Inquiry and Word Count) [10], a text analysis software program that calculates the degree to which people use various categories of words across a wide array of texts genres. They reported that pronouns and discrepant terms are good predictors of the conceptual quality of student responses. Some of our features are informed by their work.

Yoo and Kim [27] worked on predicting the project performance of students and student groups based on stepwise regression analysis on dialogue features in Online Q&A discussions. To extract dialogue features they made use of LIWC too, but also of Speech Acts [11], a tool for profiling user interactions in on-line discussions. They found that the degree of information provided by students and how early they start to discuss before the deadline are 2 important factors in explaining project grades. A similar research was conducted by Romero and colleagues [13], who also included (social) network related features. Their statistical analysis showed that the best predictors related to students' dialogue are the number of contributions (messages), the number of words, and the average score of the messages.

3. THE DATA

We conducted our research on log-files from experiments with DeepTutor [14], the first ITS based on the framework of Learning Progressions (LPs) [3]. DeepTutor is a conversational ITS based on constructivist theories of learning, which encourages students to self-explain solutions to complex science problems and only offers help, in the form of progressively informative hints, when needed. This type of adaptivity, within a task, is also known as *micro-adaptivity* [21]. Our approach to predict students' knowledge levels relies on the fact that each DeepTutor-student dialogue has its own characteristics, strongly influenced by student's profile.

Our work is based on data collected from students interacting with DeepTutor after-school, outside the lab, in the course of a multi-session online training experiment that took place in the fall of 2013. This was possible because DeepTutor is a fully-online conversational ITS, accessible from any device with an Internet connection. During the experiment, students took a pre-test, trained with the system for about one hour each week for a period of 3 consecutive weeks, and then took a post-test. Each training session consisted in solving a sequence of 8 physics problems with help from DeepTutor. The pre-test and post-test were taken under the strict supervision of a teacher. During the 3 training sessions, students were exposed to 3 different topics, one topic per week, in the following fixed order: force and motion (Newton's first and second laws), free-fall, vectors and motion (2-D motion). In our analysis, we included only 150 the students who finished all sessions in one sitting. They were randomly assigned to one of two *conditions* mentioned earlier: micro-adaptive-only (μA; n=70) and fully-adaptive (ϕA; n=80). In this paper, we only analyze the dialogues corresponding to the first session of training as it was closest to the pre-test. The data consists of a total of 8,191 student dialogue turns (9,256 sentences) out of which 4,587 (5,102 sentences) belong to μA condition, and 3,604 (4,154 sentences) to ϕA condition. Before feature extraction, the dialogues were preprocessed using Stanford NLP Parser [18]. The preprocessing pipeline consisted in 5 steps: tokenization, sentence splitting, part of speech (pos) tagging, lemmatization, and chunking.

4. THE FEATURES

The features we mined from dialogues was inspired by the work mentioned in section 2 of the paper, as well as by studies on automated essay scoring [16] in which the goal was to infer students' knowledge levels or skills from their essays. Also, our set of features is grounded in the learning literature as explained next.

The proposed dialogue features can be classified into 3 major categories: ***time-on-task***, ***generation***, and ***pedagogy***. In general, *time-on-task*, which reflects how much time students spend on a learning task, correlates positively with learning [20]. We measured *time-on-task* in several different ways as: total time (in minutes) or normalized total time (using the longest dialogue as the normalization factor). Additional time-related features were extracted such as the average time per turn and winsorized versions of the basic time-related features. *Generation* features are about the amount of text produced by students. Greater word production has been shown to be related to deeper levels of comprehensions [2, 22]. *Pedagogy* features refer to how much scaffolding a student receives (e.g. number of hints) during the training. Scaffolding is well documented to lead to more learning than lecturing or some other less interactive type of learning such as reading a textbook [22]. Feedback is an important part of scaffolding and therefore we extracted features regarding the type (positive, neutral, negative) and frequency of the feedback [17].

We extracted raw features as well as normalized versions of the features. In some cases, the normalized versions seem to be both more predictive and more interpretable. For instance, the number of hints could vary a lot from simpler/short problems, where the solution require less scaffolding in general even for low knowledge students, to more complex problems which would require more scaffolding as there are more steps in getting to the solution. That is, a normalized feature, such as the percentage of hints, would allow us to better compare the level of scaffolding in terms of hints across problems of varying complexity or solution length. In our case, we normalized the number of hints by the maximum number of hints a student may receive when answering vaguely or incorrectly at every single step during the dialogue. This number can be inferred from our dialogue management components.

We mined a total of 43 features from 1,200 units of dialogue which led to 43×1,200= 51,600 measurements. The unit of dialogue analysis was a single problem in a training session. Because the force-and-motion training session consisted of 8 problems, and we collected 150 sessions from 150 students, we ended up with 8×150=1,200 units. Due to space constraints, we do not provide the full list of features:

Time-on-task features: *total_time* (the time length of the dialogue in minutes), *avg_time_per_turn* (the average length of a student turn in minutes);

Generation features: *dialogue_size* (length of the student dialogue (number of words, no punctuation included)), *avg_dialogue_size_per_turn*, *#sentences* (number of sentences), *#chunks* (number of syntactic constituents), *vocSize* (vocabulary size), *content_vocSize* (vocabulary size of content words), *non_content_vocSize*, *dialogue_length_div_voc* (#words divided to vocabulary size), *%physicsTerms* (percentage of physics related words out of those used), *%longWords* (percentage of words longer than 6 characters), *posDiversity* (number of unique different pos-es divided by vocabulary size), *%puctuation* (percentage of punctuation out of all tokens), *%articles*, *%pronouns*, *%self-references*, *totalIC* (total Information Content of the dialogue: explained below), *totalIC_per_word*, *positiveness* (text positiveness computed based on SentiWordNet: explained below), *negativeness*.

Scaffolding features: *#turns* (number of student turns), *#normalized_turns*, *#c_turns* (number of student turns classified as contributions), *%pos_fb* (percentage of turns for which student received positive feedback), *%neg_fb*, *pos_div_pos+neg* (positive feedback divided by positive plus negative feedback), *vague_div_vague+pos* (neutral feedback divided by positive plus neutral feedback), *#shownHints* (number of shown hints), *#shownPrompts* (number of shown prompts), *#shownPumps* (number of shown pumps).

Next, we discuss on short the dialogue features on the Information Content and positive-negative polarity.

Information Content (IC) was used by Resnik [12] to measure the informativeness of a *concept c*, on the assumption that the lower the frequency of *c*, the higher its informativeness. Resnik made use of Princeton WordNet [5] and its hierarchical taxonomy, where each node is a concept, also called *synset*. The more general concepts are at the top of the hierarchy, while the specific ones, at the bottom. Each synset can be realized in texts by any of the specific senses of certain words (i.e. *literals*), which are considered to be part of that synset. To count the frequency of a nominal synset *s* in a reference text, Resnik sums the frequencies of the literals of *s* and those of all the synsets for which *s* is a parent in the hierarchy. Thus, the estimated probability of occurrence can be easily computed and so is the IC value for that synset.

We replicated Resnik's work on WordNet 3.0 for all pos-es. Starting from synsets, we transferred the IC values to individual words. If a word has various senses associated with different synsets, we assign to it the IC value corresponding to the most non-informative synset, so that high IC values are only associated with informative words. For a word that does not appear in WordNet, our algorithm selects a WordNet literal of the same grammatical category, so that the similarity between the two is sufficiently high according to an LSA model built on the whole Wikipedia [19]. We compute the IC of a text as the sum of the IC values for its words.

To include features on the ***Positive-Negative Polarity*** of the dialogues, we made use of an updated version [1] of *SentiWordNet* [4] in which, each WordNet synset is assigned scores representing the polarity strength on 3 dimensions: Positive, Negative and Objective. Based on SentiWordNet, we extracted two lists: one of positive and the other one of negative words along with computed scores for positivity and respectively, negativity. For each word in WordNet, we summed up the values on all 3 polarity dimensions corresponding to the synsets that contain that word. If the Positive dimension value (p) is at least twice the Negative value (n) and also p is greater or equal to the Objective value (o) or greater than a certain threshold (> 2), than that word is added to the list of positive words with a positivity value computed as p / (p+ n + o). An identical procedure is applied for finding the negative words. Dialogue positiveness (negativeness) is simply computed as the sum of the values assigned to all positive (negative) words found in the text, divided by the total number of words in that text.

5. EXPERIMENTS AND RESULTS

Our larger goal is to understand how various dialogue units, corresponding to one problem in a session, individually and as groups, relate to students' prior knowledge as measured by the pre-test, which is deemed as an accurate estimate of students' knowledge level. The group analysis would indicate after how much dialogue, corresponding to consecutive training problems, one can accurately infer students' pre-test score. We present in this paper only our initial feature analysis, due to space reasons.

5.1 Feature Analysis

We started by extracting all the above-mentioned features for the sub-dialogues corresponding to individual problems. We worked on 1,193 sub-dialogues spanning over 7,927 turns (6.64 on average), 5,441 minutes (4.56 on average), with a total length of 74,036 words (62.05 on average). The next step was to identify the features whose values best correlate with the pre-test scores. We considered both the entire pre-test (an extended version of Force Concept Inventory) [8], which can be seen as assessing students overall knowledge with respect to Newtonian Physics, but also the *pre-testFM*: the portion of the pre-test containing questions directly related to the force-and-motion training session. Table 1 shows correlations of features with the pre-test scores for μA condition.

Table 1. Correlations values with pre-test (top) and pre-testFM (bottom) for interesting features on each of the 8 problems in the μA condition.

	1	2	3	4	5	6	7	8
f1	-0.36	-0.408	-0.141	-0.176	-0.225	-0.136	-0.254	-0.181
	-0.408	-0.333	-0.162	-0.182	-0.256	-0.225	-0.25	-0.219
f2	0.344	0.262	0.242	0.202	0.213	0.23	0.321	0.236
	0.358	0.221	0.254	0.183	0.157	0.125	0.267	0.216
f3	-0.423	-0.403	-0.303	-0.295	-0.35	-0.245	-0.283	-0.225
	-0.433	-0.293	-0.268	-0.296	-0.305	-0.258	-0.29	-0.228
f4	-0.448	-0.444	-0.333	-0.308	-0.34	-0.36	-0.361	-0.276
	-0.473	-0.334	-0.305	-0.295	-0.278	-0.351	-0.331	-0.254
f5	0.458	0.368	0.193	0.36	0.254	0.208	0.311	0.264
	0.458	0.297	0.122	0.386	0.206	0.168	0.251	0.23
f6	-0.424	-0.425	-0.215	-0.291	-0.284	-0.326	-0.415	-0.326
	-0.464	-0.314	-0.248	-0.318	-0.223	-0.317	-0.393	-0.29
f7	-0.404	-0.386	-0.295	-0.352	-0.28	-0.337	-0.194	-0.2
	-0.385	-0.284	-0.225	-0.31	-0.219	-0.304	-0.158	-0.208

Table 1 shows that with some exceptions for problem 5, the time length (**f1**), the number of sentences (**f3**), the number of turns (**f4**), and the number of hints (**f6**) and prompts shown (**f7**) have negative correlations with the pre-test scores, while the average word-length of a turn (**f2**) and the percentage of turns receiving positive feedback (**f5**) have positive correlations. These outcomes confirm similar findings from previous studies [22]. Interestingly enough, the number of sentences students produce seem to be less and less correlated with the pre-test scores as the students advance through the training session.

Correlations for the ϕA condition were derived using a more complex process given that students were grouped into 4 knowledge levels based on their overall pre-test score and so, 4 different sets of problems were used. As such, we could not conflate the data across knowledge groups and therefore we studied the correlations for each set of problems separately. In this case, because of macro-adaptation, but also because the number of dialogues for each knowledge level was much smaller (80 students

were grouped in 4 knowledge level groups: low (14), medium low (15), medium high (21), and high (30)), the best correlated features were somehow different. Given the space constraints, we will present these results in a future paper.

6. CONCLUSIONS AND FUTURE WORK

This paper presented our work towards predicting students' prior knowledge based on the characteristics of their dialogue while engaging in problem solving with a conversational ITS. The proposed dialogue features can be classified into three major categories: time-on-task, generation, and pedagogy. The features were analyzed throughout an entire training session using instructional tasks as the unit of analysis. Our next step would be to analyze these features across increasing subsets of instructional tasks, e.g. the first Physics problem in a session vs. first two problems vs. first three problems, in order to investigate after how many instructional tasks the features best correlate with prior knowledge. It should be noted that the more tasks into a session we consider the more likely the student model may have significantly change, due to training, compared. Furthermore, we will investigate these features for students at different prior knowledge level, e.g. low knowledge vs. high knowledge students. Finally, we plan to investigate prediction models based on the analyzed features and also to add affect-related features.

7. ACKNOWLEDGMENTS

This research was supported in part by Institute for Education Sciences under awards R305A100875. Any opinions, findings or recommendations expressed in this material are solely the authors'.

8. REFERENCES

[1] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 2200-2204).

[2] Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., and Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, 25, 471–533.

[3] Corcoran, T., Mosher, F.A., and Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform.* Consortium for Policy Research in Education Report #RR-63. Philadelphia, PA.

[4] Esuli, A., and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC* (Vol. 6, pp. 417-422).

[5] Fellbaum, C. (1998). WordNet: An electronic lexical database.

[6] Forbes-Riley, K., and Litman, D. J. (2006). Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters. In *Proceedings of HLT Conference*.

[7] Forbes-Riley, K., Litman, D., Purandare, A., Rotaru, M., and Tetreault, J. (2007). Comparing Linguistic Features for Modeling Learning in Computer Dialogue Tutoring. In Proc. of the 13th International Conference on AIED, LA, CA.

[8] Hestenes, D., Wells, M., and Swackhamer, G. (1992). Force concept inventory. The physics teacher, 30(3), 141-158.

[9] Lintean, M., Rus, V., and Azevedo, R. (2011). Automatic Detection of Student Mental Models during Prior Knowledge Activation in MetaTutor, *International Journal of Artificial Intelligence in Education*, 21(3), 169-190.

[10] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71.

[11] Ravi, S., and Kim, J. (2007). Profiling student interactions in threaded discussions with speech act classifiers. *Frontiers in Artificial Intelligence and Applications*, 158, 357.

[12] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007.

[13] Romero, C., López, M. I., Luna, J. M., and Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, 68.

[14] Rus, V., D'Mello, S., Hu, X., and Graesser, A. C. (2013). Recent Advances in Conversational Intelligent Tutoring Systems. AI Magazine, 34(3). *Lang. Syst.* 15, 5, 795-825.

[15] Rus, V., Stefanescu, D., Baggett, W., Niraula, N., Franceschetti, D., & Graesser, A.C. (2014). Macro-adaptation in Conversational Intelligent Tutoring Matters, *The 12th International Conference on Intelligent Tutoring Systems*, June 5-9, Honolulu, Hawaii.

[16] Shermis, M. D., and Burstein, J. (2003). Automated Essay Scoring: A Cross Disciplinary Perspective. Mahwah, NJ: Lawrence Erlbaum Associates.

[17] Shute, V. J. (2008). Focus on Formative Feedback. Review of Educational Research, 78(1), 153 -189.

[18] Socher, R., Bauer, J., Manning, C.D., and Ng, A.Y. (2013). Parsing With Compositional Vector Grammars. Proceedings of ACL 2013.

[19] Stefanescu, D., Banjade, R., and Rus, V. (2014). Latent Semantic Analysis Models on Wikipedia and TASA, LREC

[20] Taraban, R., and Rynearson, K. (1998). Computer-based comprehension research in a content area. Journal of Developmental Education, 21, 10-18.

[21] VanLehn, K. (2006). The behavior of tutoring systems. International Journal of AI in Education. 16 (3), 227-265.

[22] VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A., and Rose, C.P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31*, 3-62.

[23] VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems, *Educational Psychologist*, 46:4, 197-221.

[24] Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the 8th conference on EACL* (pp. 271-280).

[25] Williams, C., and D'Mello, S. (2010). Predicting student knowledge level from domain-independent function and content words. In *Intelligent Tutoring Systems* (pp. 62-71).

[26] Woolf, B. (2008). Building Intelligent Interactive Tutors, Student-Centered Strategies for Revolutionizing E-Learning, Elsevier & Morgan Kaufmann Publishers, 2008.

[27] Yoo, J.-B., and Kim, J. (2012). Predicting Learner's Project Performance with Dialogue Features in Online Q&A Discussions. In *Intelligent Tutoring Systems* (pp. 570-575).