Assessing student contributions in a simulated human tutor with Latent Semantic Analysis

Peter Wiemer-Hastings (PWMRHSTN@MEMPHIS.EDU)
Katja Wiemer-Hastings (KWIEMER@CC.MEMPHIS.EDU)
Arthur C. Graesser (A-GRAESSER@MEMPHIS.EDU)
Department of Psychology; Campus Box 526400
Memphis, TN 38152-6400 USA

Introduction

One-on-one tutoring is a highly-effective means of education compared to classroom instruction. But what accounts for its learning gains? We are developing an intelligent tutoring system called AutoTutor which is based on studies of human tutors. This paper describes the findings from these studies that form the foundation of our tutor. We also show how a corpus-based, statistical natural language understanding technique called Latent Semantic Analysis (LSA) allows AutoTutor to understand student responses and respond appropriately. Finally, we describe analyses of the performance of LSA with respect to human raters.

Psychological foundations of AutoTutor

A fairly complete description of the architecture of AutoTutor has been given elsewhere [Wiemer-Hastings et al., 1999]. Here, we give a brief description of the tutorial discourse foundations of the system in order to highlight how LSA is used by it.

Graesser et al (1997) compared the frequency of a number of features between one-on-one tutoring and classroom education. They found that the following types of activities occur rarely, or at least not more often than in the classroom: active student learning, convergence toward shared meanings, error diagnosis, anchored learning, and sophisticated pedagogical strategies. The following types of activities were significantly more prevalent in tutoring situations than in classroom teaching: use of examples, curriculum scripts, explanatory reasoning by the student and tutor, and collaborative question answering and problem solving.

Assessing student contributions with LSA

LSA relies on a statistical technique that reduces the co-occurrence information in a corpus to a high-dimensional space in which meanings of words and sentences are represented as vectors. The similarity between any two meaning vectors can be computed by calculating the cosine between them. Previous work has shown human-like performance by LSA on variety of tasks.

We trained LSA on two textbooks, 30 articles, and the items of our curriculum script: the questions, expected good answers, and responses that AutoTutor uses. We used a 200-dimensional LSA space to compare student contributions with expected good answers, and calculated a compatibility score that reflected the extent to which the student contribution matched part of the good answer.

We tested LSA's performance by comparing its ratings to those of four human raters: two subject-area experts, and two with intermediate domain knowledge. The correlation between the two intermediate-knowledge raters was r=0.52. The correlation between the two expert raters was r=0.78. The correlation between the average human rating and the LSA rating was r=0.47, almost equalling the interrater reliability for the intermediate-knowledge human raters.

Effects of student text attributes on LSA

We performed an ANOVA with the LSA compatibility scores as the dependent variable, and these three independent variables measuring attributes of the student contributions: (1) *quality*: measured by the compatibility score from the human raters, (2) *length*: the number of words in the student speech contribution, and (3) *information content*: the number of glossary terms in the contribution divided by the number of words. As expected, there was a main effect of the quality of the student contributions. If the absolute LSA score is caused simply by longer contributions, there would be a main effect of the number of words. There was not a significant effect here however. There was a main effect of the information content of the student contributions. If the density of glossary terms in the student contributions is indicative of their quality, this suggests that LSA is measuring the right thing. However, because this effect was independent of the effect of the human quality judgment, it also suggests that there is something to the LSA judgments which is independent of the quality of the contribution alone, at least as human raters judge it.

Acknowledgments

This project is supported by grant number SBR 9720314 from the National Science Foundation's Learning and Intelligent Systems Unit.

References

[Graesser et al., 1997] Graesser, A., Millis, K., and Zwaan, R. (1997). Discourse comprehension. In Spence, J., Darley, J., and Foss, D., editors, *Annual Review of Psychology*, volume 48. Annual Reviews Inc., Palo Alto, CA.

[Wiemer-Hastings et al., 1999] Wiemer-Hastings, P., Wiemer-Hastings, K., and Graesser, A. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In *Proceedings of Artificial Intelligence in Education*, Amsterdam. IOS Press.