Making AutoTutor Agents Smarter: AutoTutor Answer Clustering and Iterative Script Authoring

Zhiqiang Cai[1(✉)], Yan Gong[1], Qizhi Qiu[2], Xiangen Hu[1], and Art Graesser[1]

[1] University of Memphis, Memphis, TN, USA
{zcai,ygong2,xhu,a-graesser}@memphis.edu
[2] Wuhan University of Technology, Wuhan, Hubei, China
qiuirene@163.com

Abstract. AutoTutor uses conversational intelligent agents in learning environments. One of the major challenges in developing AutoTutor applications is to assess students' natural language answers to AutoTutor questions. We investigated an AutoTutor dataset with 3358 student answers to 49 AutoTutor questions. In comparisons with human ratings, we found that semantic matching works well for some questions but poor for others. This variation can be predicted by a measure called "question uncertainty", an entropy value on semantic cluster probabilities. Based on these findings, we propose an iterative AutoTutor script authoring process that can make AutoTutor agents smarter and improve assessment models by iteratively adding and modifying both questions and ideal answers.

Keywords: AutoTutor · Conversational agents · ASAT · ITS · Authoring tool · Assessment · Short answer grading · Question uncertainty

1 Introduction

The AutoTutor project was launched in 1997 by Art Graesser and his colleagues. Since then, many AutoTutor applications have developed in different domains, including physics, computer literacy, psychology, algebra, and electronics. Researches have reported that AutoTutor learning gains are estimated to be between 0.3 and 0.8 sigma [1].

AutoTutor agents use scripted content that is intelligently selected during tutoring. The major discourse mechanism is called Expectation-Misconception-Tailored (EMT) dialogue [2]. AutoTutor starts tutoring by introducing a problem and asking a main question. If a learner cannot answer the main question well, then a sequence of hints and prompts are asked to help the learner improve the answer. A hint is a question that attempts to elicit an answer of approximately a clause or sentence. A prompt is a question targeting a specific concept, usually a word or a phrase. When the learner expresses a misconception, AutoTutor corrects the misconception and then continue with hints, prompts, and assertions until all expectations are covered.

Evaluating student answers to AutoTutor questions is a task known as "short answer grading" [3]. Cai et al. [4] reported that LSA (Latent Semantic Analysis) [5] could play

© Springer International Publishing AG 2016
D. Traum et al. (Eds.): IVA 2016, LNAI 10011, pp. 438–441, 2016.
DOI: 10.1007/978-3-319-47665-0_50

an important role in grading such short answers in addition to regular expressions. This paper further investigates the use of LSA and proposes an iterative script authoring process to improve grading accuracy.

2 An AutoTutor Dataset with Human Ratings

The AutoTutor dataset in this paper was extracted from log files in three experiments conducted in spring 2002, fall 2002 and summer 2003. The subjects of the experiments were college students. Each student interacted with a subset of the 10 problems in conceptual physics. A total of 512 log files were received, containing 7584 student responses to 247 AutoTutor questions. The student answers were rated with 1–6 scale by two experts (both are co-authors); one is a professor in computer engineering and the other is a graduate student in computer engineering. 120 student answers were randomly sampled from the 7584 answers to train the two raters. Two raters first rated the 120 training answers independently. Then they sat together and went through the items with rating differences greater than 2. They achieved agreement and then rated the rest of the 7464 items. The correlation between the two raters' scores was 0.828. They rated 92.8 % of the answers with a difference not greater than 2. In this analysis we selected a subset of the answers satisfying: (1) the difference between the two raters' ratings is less than 2; and (2) the associated question has at least 50 answers. The selected subset contained 38 hints, 11 prompts, and 3358 student answers. In the rest of this article, we will use "human rating" to refer to the average rating of the two human ratings.

3 Performance of Semantic Matching

For each of the 3358 responses, we computed the LSA cosine between the student answers and the hint/prompt answers given by script authors. We then computed the correlation between LSA cosines and human ratings in each question. It turned out that the correlations were very different, ranging from −0.174 to 0.995. The correlations for prompts were high; 7 out of the 11 selected prompts had correlations higher than 0.7, but there was one below 0.3. The correlations on hints were much lower. Only 7 out of 38 were above 0.7, 11 out of 38 were below 0.3, and 4 of them were even below 0.

Why did LSA work well for some questions but poor for others? There could be multiple reasons. The major reason could be that some questions have "convergent" answers. That is, answers of such questions are semantically similar. Other questions could have "uncertain" answers. That is, the answers may be in multiple semantic groups. To investigate this, in the next section, we propose a "question uncertainty" measure based on semantic clustering and then show that question uncertainty predicts LSA performance on questions.

4 Question Uncertainty

For a given question Q, suppose the collected student answers are grouped into N clusters according to a particular semantic threshold t. We define the uncertainty of Q with threshold t as the "normalized" entropy value of the clustering of the responses:

$$U(Q,t) = -\frac{\sum_{i=1}^{N} p_i log p_i}{log N} \tag{1}$$

p_i is the number of responses in cluster i divided by the total number of responses. The uncertainty is normalized so the value ranges from 0 to 1.

The above definition depends on the semantic clustering. There are many different types clustering algorithms. We used the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm. This algorithm automatically and reasonably determines the number of clusters. The clustering depends on three choices: (1) distance function; (2) distance threshold; and (3) minimum number of elements in a cluster. The formula ($1 - LSA\ cosine$) was used as distance between any two responses. LSA cosine values are typically positive. Occasionally negative cosine values appear. To make the distance value fall into the [0, 1] interval, negative cosine values were set to zero. Clustering analyses were performed for threshold t from 0.05 up to 0.95 with a 0.05 increase in each increment. The minimum number of elements was set to 2 so that the clustering result could contain any number of elements, with single element clusters as outliers. For each threshold $t(t = 0.05, 0.1, \ldots, 0.95)$, the correlation between the question uncertainty and the LSA performance (i.e., correlation between LSA cosine and human rating) was computed. The correlation as a function of the semantic distance threshold t resembled a cubic function. It started with negative correlation at $t = 0.05$ and decreased to the deepest point of the valley (-0.51) at $t = 0.25$. It then monotonically increased to the top point of the hill (0.50) at $t = 0.90$.

The negative correlation at $t = 0.25$ indicates that when the threshold is set to 0.25, question uncertainty can predict LSA performance. That is, for an appropriate threshold, if most answers fall into a small number of clusters, LSA performance is high. Otherwise, if the answers evenly fall into many clusters, LSA performance is low.

The high positive correlation at $t = 0.90$ is interesting. When the distance threshold is so high, the answers should usually fall into a single cluster, resulting in zero uncertainty. However, if the answers are still divided into multiple clusters with such a high distance threshold, that means there are multiple groups of answers that are semantically very different and thus can be easily classified by LSA.

5 Iterative AutoTutor Script Authoring

Based on the above discussions, LSA does not always perform well in AutoTutor answer assessment. The reason is that AutoTutor questions do not always have answers in a single semantic group. In recent AutoTutor systems, we already allow authors to create

multiple good answers. However, it is really hard for authors to imagine all possible semantic groups student answers may form. Therefore, an iterative authoring process is inevitable.

An iterative authoring process starts with an initial script that contains authored answers to questions. AutoTutor agents may sometimes generate incorrect feedback and subsequent hints/prompts at the beginning because of poor semantic performance. After enough student responses are collected, however, the AutoTutor answer analysis model would cluster student answers and provide question uncertainty scores to authors. For questions with high uncertainty at a lower semantic distance threshold, authors are informed to review the student responses. Based on how the student responses cluster, authors may improve AutoTutor scripts by: (a) revising existing questions and answers; (b) adding new answers to a question; and (c) adding or revising contextual feedback to each response cluster. Once new script questions and answers are provided, AutoTutor uses new answer clusters to assess student answers, which provides more accurate answer classification, feedback, hints, and prompts.

Acknowledgement. This research was supported by the National Science Foundation (SBR 9720314, REC 0106965, REC 0126265, ITR 0325428, REESE 0633918, ALT-0834847, DRK-12-0918409, 1108845), the Institute of Education Sciences (R305H050169, R305B070349, R305A080589, R305A080594, R305G020018, R305C120001), Army Research Lab (W911INF-12-2-0030), and the Office of Naval Research (N00014-00-1-0600, N00014-12-C-0643, N000014-16-C-3027). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, IES, or DoD.

References

1. Nye, B.D., Graesser, A.C., Hu, X.: AutoTutor and family: a review of 17 years of natural language tutoring. Int. J. Artif. Intell. Educ. **24**, 427–469 (2014)
2. Graesser, A.C.: Conversations with AutoTutor help students learn. Int. J. Artif. Intell. Educ. **26**, 124–132 (2016)
3. Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. Int. J. Artif. Intell. Educ. **25**(1), 60–117 (2015). Springer
4. Cai, Z., Graesser, A.C., Forsyth, C., Burkett, C., Millis, K., Wallace, P., Halpern, D., Butler, H.: Trialog in ARIES: user input assessment in an intelligent tutoring system. In: Chen, W., Li, S. (eds.) Proceedings of the 3rd IEEE International Conference on Intelligent Computing and Intelligent Systems, pp. 429–433. IEEE Press, Guangzhou (2011)
5. Landauer, T., McNamara, D.S., Dennis, S., Kintsch, W. (eds.): Handbook of Latent Semantic Analysis. Erlbaum, Mahwah (2007)