

Expecting the Unexpected: Warehousing and Analyzing Data from ITS Field Use

W. Lewis Johnson, Naveen Ashish, Stephen Bodnar, Alicia Sagae

Alelo Inc, 12910 Culver Bl., Suite J, Los Angeles, CA 90066 USA
{ljohnson, nashish, sbodnar, asagae}@Alelo.com

Abstract. One should expect the unexpected when deploying intelligent tutoring systems. This paper describes a case study in collecting, warehousing, and analyzing field usage data from two language and culture learning environments, to understand what happened when they were deployed. A data warehousing system, Hoahu, was used to process the raw data and transform it into a relational database to facilitate queries and analysis. The system also supported data annotation by subject matter experts to facilitate comparison of automated assessments against human raters. Errors and inconsistencies in the data were identified and corrected. The resulting data warehouse has proven valuable for understanding the trajectory of learning over extended periods of time and analyzing the strengths and weaknesses of complex interactive subsystems such as spoken dialog systems.

Keywords: empirical studies, educational data mining, language learning, dialog systems

1 Introduction

It is often difficult to predict how intelligent tutoring systems (ITSs) perform in the field. Patterns of learner performance may emerge over time that were not apparent in short-duration tests common to formative evaluations. The performance of the system depends in part on the learning trajectory of each learner. Such issues arise quite commonly with the interactive learning environments that Alelo develops, which learners employ for tens or even hundreds of hours, and which incorporate animated characters that engage in many spoken conversations with learners.

This paper presents a case study in which data from field use of two intelligent tutoring systems were collected, warehoused, and analyzed. Prior to the release of the latest versions of our Iraqi Arabic and Sub-Saharan French courses, Naval personnel at several sites around the United States volunteered to take the courses in self-study mode, in their spare time. After the trainees completed their training, we retrieved logs and speech recordings from the training sessions. We then used a data warehousing system, named *Hoahu* to process and organize the data in a database for analysis and query purposes. (*Hoahu* means “to collect” in Hawaiian.)

2 Hoahu Data Warehouse

A schematic overview of Hoahu is provided in Fig. 1. One can look at Hoahu as a pipeline that takes raw log data and recordings and transforms them to a form amenable for high-end analysis. The pipeline works as follows. Data in the logs are first sent through *Kapaa*, the anonymizer module in Hoahu, which creates an anonymized data image that we then elaborate on. The data are then processed by *Ono*, a module that identifies and extracts objects of interest.

Ono creates a relational representation of these objects, and identifies and stores relationships between objects, specifically containment relationships. Analysts can then use this database for analysis of log data – at present we are using structured queries (SQL) over the database.

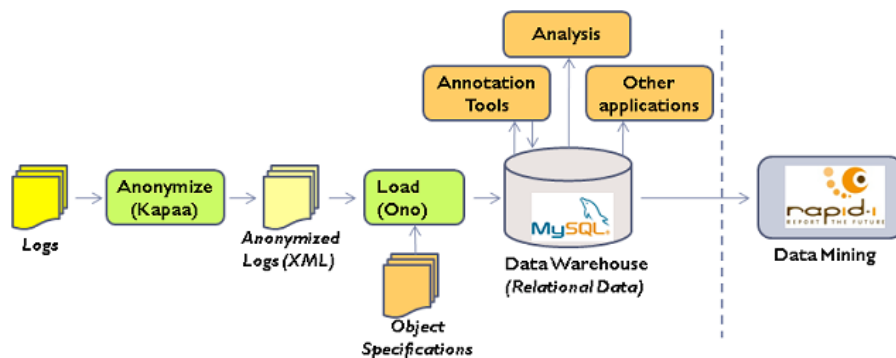


Fig. 1. Hoahu Architecture.

Mostow & Beck [1] present a good summary of research activities that can be supported with appropriate system logging ("why to log"), including reporting, mining, and browsing historical data, in addition to tutoring itself. They also present a set of recommendations for *what* and *how* to log. Hoahu adapts these recommendations to a large-scale application context, improving on prior work where logs were analyzed using ad-hoc tools [2]. Other systems that are consistent with the Mostow & Beck model include DataShop [3] and the ASSISTment Builder [4].

3 Results

The following analysis focuses on trainees who completed at least 4 hours of training (8 out of 25 Arabic trainees and 5 out of 20 French trainees). We were particularly interested in studying *dialog breakdowns*, or what Jordan, et. al, [5] call *mis-hearings* and *misunderstandings*. We define a dialog breakdown as any dialog turn where one speaker (e.g., the learner) says something and the other speaker (e.g., an agent) responds in a way that suggests that it did not understand. A certain number of dialog

breakdowns is anticipated, just as breakdowns occur in real life when language learners interact with native speakers. However persistent dialog breakdowns are likely to lead to learner frustration.

We chose $N \geq 4$ as the threshold for the number of utterance attempts at which point the dialog breakdown was considered unacceptably severe. The breakdown rates for the two languages were very similar, 6.5% for Arabic vs. 7.1% for French. They were substantially lower than the rate in an earlier pilot study with intermediate French speakers (18.6%). When focusing on the exercises in common between the two data sets, the rate was also lower (7.9% vs. 18.6%). The mean number of utterance attempts per dialog turn was 1.73 in the Arabic data and 1.64 in the French data, vs. 2.2 in the pilot data. Reviews of the speech recordings of the different groups revealed that the speech of the field trial learners was very different from that of the pilot test learners in terms of complexity and pronunciation accuracy. Many of the beginning learners' utterances are very badly pronounced, or even unintelligible. This illustrates the importance of using authentic field data to assess system performance.

We are currently having human raters annotate samples of the learner data, to judge the intelligibility of the speech and the accuracy of the system's interpretation of the speech. In contrast with typical speech recognition applications, our goal is *not* to achieve the highest possible speech recognition rates, but rather to correctly recognize intelligible speech and to reject and diagnose errors in errorful speech.

Meaningful response rates, when the agents understood and responded to the learners' speech, were 58% for Arabic and 59% for French. When learners did not rely on hints, the rates were 51% for Arabic and 58% for French. The speech recognition acceptance rates were 66% for Arabic and 57% for French. Learners must repeatedly attempt utterances until they get a meaningful response, which tends to multiply the number of recognition failures and non-meaningful responses. The gap between recognition and response rates can result from issues with the dialog model.

Acknowledgments. This work was generously supported by the Office of Naval Research under the ISLET project.

References

1. Jack Mostow, J., Beck, J.: What, How, and Why should Tutors Log? In: Proceedings of EDM 2009, pp. 269-278 (2009)
2. Johnson, W.L., Wu, S.: Assessing aptitude for learning with a serious game for foreign language and culture. In: Proceedings of ITS 2008, pp. 520-529. Springer, Berlin (2008)
3. Koedinger, K., Cunningham, K., Skogsholm A., Leber, B.: An open repository and analysis tools for fine-grained, longitudinal learner data. In: Proceedings of EDM 2008, pp. 157-166 (2008)
4. Razzaq, L., Patvarczki, J., Almeida, S.F., Vartak, M., Feng, M., Heffernan, N.T. and Koedinger, K.: The ASSISTment builder: Supporting the Life-cycle of ITS Content Creation. IEEE Transactions on Learning Technologies Special Issue on Real-World Applications of Intelligent Tutoring Systems. 2(2), pp. 157-166 (2009)
5. Jordan, P., Litman, D., Lipschultz, M., Drummond, J.: Evidence of Misunderstandings in Tutorial Dialogue and their Impact on Learning. In: Artificial Intelligence in Education, pp. 125-132. IOS Press, Amsterdam (2009)