Predicting Dialogue Breakdown in Conversational Pedagogical Agents with Multimodal LSTMs

Wookhee Min[1], Kyungjin Park[1], Joseph Wiggins[2], Bradford Mott[1],
Eric Wiebe[1], Kristy Elizabeth Boyer[2], and James Lester[1]

[1] Center for Educational Informatics, North Carolina State University, Raleigh NC 27606, USA
{wmin, kpark8, bwmott, wiebe, lester}@ncsu.edu
[2] Department of Computer & Information Science & Engineering, University of Florida,
Gainsville, FL 32601, USA
{jbwiggi3, keboyer}@ufl.edu

Abstract. Recent years have seen a growing interest in conversational pedagogical agents. However, creating robust dialogue managers for conversational pedagogical agents poses significant challenges. Agents' misunderstandings and inappropriate responses may cause breakdowns in conversational flow, lead to breaches of trust in agent-student relationships, and negatively impact student learning. Dialogue breakdown detection (DBD) is the task of predicting whether an agent's utterance will cause a breakdown in an ongoing conversation. A robust DBD framework can support enhanced user experiences by choosing more appropriate responses, while also offering a method to conduct error analyses and improve dialogue managers. This paper presents a multimodal deep learning-based DBD framework to predict breakdowns in student-agent conversations. We investigate this framework with dialogues between middle school students and a conversational pedagogical agent in a game-based learning environment. Results from a study with 92 middle school students demonstrate that multimodal long short-term memory network (LSTM)-based dialogue breakdown detectors incorporating eye gaze features achieve high predictive accuracies and recall rates, suggesting that multimodal detectors can play an important role in designing conversational pedagogical agents that effectively engage students in dialogue.

Keywords: Conversational Pedagogical Agent, Dialogue Breakdown Detection, Multimodal, Natural Language Processing, Gaze.

1 Introduction

Recent years have seen the emergence of increasingly robust conversational agents paralleling significant advances in natural language processing [1]. A particularly important line of research on conversational agents investigates conversational pedagogical agents [2, 3]. They have demonstrated significant potential in intelligent tutoring systems as an effective approach to engaging students in tutorial dialogue [4], assessing student knowledge [5], and supporting learning [6]. Conversational pedagogical agents can play a central role in student interactions in game-based learning environments by enhancing students' engagement and facilitating learning through customized narratives and adaptive problem-solving support [7–9].

It is critical that conversational pedagogical agents effectively prevent dialogue breakdown, which is a conversational phenomenon in which a dialogue cannot easily proceed [10]. Dialogue breakdown occurs when an agent misunderstands what a human intends to communicate and, as a result, responds inappropriately. A robust dialogue breakdown detection (DBD) framework could inform conversational pedagogical agents of the need to adaptively modify their dialogue strategies to prevent breakdowns and implement a dialogue recovery strategy [11], and also could enable researchers to examine causes of breakdown in the context of error analysis [12].

In this paper, with the objective of preemptively preventing dialogue breakdown, we investigate multimodal data streams to model human dialogue behaviors. Specifically, we examine four channels: natural language utterances, eye gaze traces, student gender, and task states. Gaze behaviors have been found to be related to cognitive [13] and affective [14] processes, and temporal patterns in eye movements are associated with humans' attention and engagement [15], boredom [14], and intention [16]. We hypothesize that these multimodal features will serve as strong predictors of DBD.

We present a multimodal DBD framework using long short-term memory networks (LSTMs) [17]. We examine 92 middle school students' interaction data with a conversational pedagogical agent in a game-based learning environment for science education [18]. We compare the LSTM-based DBD framework's predictive performance to linear chain conditional random fields (CRFs) as well as support vector machines (SVMs).

2 Dialogue Breakdown Detection in CRYSTAL ISLAND

CRYSTAL ISLAND is a game-based learning environment for middle school microbiology [18]. As an extension of the game-based learning environment, we incorporated a conversational pedagogical agent within the game to investigate both affective and cognitive influences on students' learning processes. We developed a state machine-based dialogue manager for this virtual agent, Alisha. Alisha's dialogue moves are made at the agent's initiative or responding to a student dialogue move. Alisha-initiated dialogue moves are triggered by student behaviors in the game, and Alisha-response dialogue moves are made in response to students' dialogue acts [19, 20].

Students played CRYSTAL ISLAND for up to three consecutive days of classroom periods or until they completed the game. Each day, they continued the game from where they ended in their prior session. We annotated dialogue data from 92 students who completed consent forms, conversed with Alisha during the study, and completed all of their surveys. Of these students, 38 identified as Female, 32 as Male, and 22 students did not report their gender. The mean age was 13.4 years (SD = 0.69).

We defined a binary annotation scheme, *no breakdown* and *breakdown*, adapted from the labels defined in the Dialogue Breakdown Detection Challenge [11]. Two human annotators labeled the dialogue corpus. Both annotators labeled approximately 20% of the entire corpus in common, achieving an inter-rater agreement of 0.765 (i.e.,

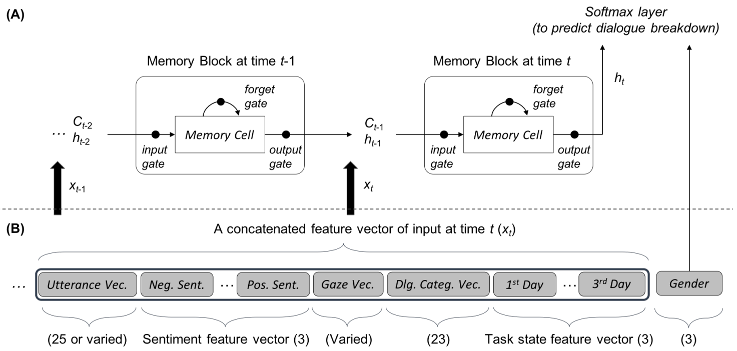

Fig. 1. (A) The LSTM-based dialogue breakdown detector. (B) An illustration of how the input at each time step is encoded. Each value in the parenthesis denotes the number of features.

substantial agreement) using Cohen's kappa [21]. In summary, the number of *breakdown* and *no breakdown* instances are 282 (23.9%) and 897 (76.1%), respectively, from the 1,179 Alisha utterances that appear in the corpus.

We adopt LSTMs (Fig. 1) to model multimodal data streams for DBD. First, we investigate linguistic features. We compare two approaches: GloVe pre-trained word embeddings [22] and a bag-of-words method. In addition, we adopt an off-the-shelf sentiment analysis toolkit [23] to identify if the current conversation is flowing in a positive or negative manner, and sentiments of student dialogues serve as an explanatory variable for DBD. Second, we use traces of objects within the game world that students were looking at in CRYSTAL ISLAND [16]. Third, we use as predictive features the history of previous Alisha dialogue move categories. Fourth, we utilize students' gender as a variable for predictive models since we have observed that female students are more considerate to Alisha than male students, who often experience more breakdown. Finally, we use task states that encode the number of gameplay sessions the student has completed. In addition, we explore an automated post-processing method, which is inspired by work in text normalization [24], to refine model predictions of breakdown in a post-hoc manner.

3 Evaluation

We evaluate model performance using student-level ten-fold cross-validation. While predictive accuracy is an important metric, recall is particularly important in this work since the primary objective is to identify potential dialogue breakdown situations in advance and adapt the current policy to avoid them. Because the corpus has an imbalanced distribution in data (only 23.9% were labeled positive, i.e., *breakdown* instances), in each fold we randomly up-sample positive examples from the training set to have a 50-50 distribution between the two labels, and evaluate trained models with the test set for which no up-sampling was applied.

In this work, we investigate two baseline models, including linear CRFs and SVMs with a radial basis function. We evaluate the models' predictive accuracy across the three machine learning techniques. A different set of hyperparameters for each of the

Table 1. Average accuracy rates over test examples in CV (**P** and **BoW** denote the post-processing technique applied and the bag-of-words method, respectively).

	Gaze+P	Gaze	NoGaze+P	NoGaze
LSTM (BoW)	**79.56**	78.29	79.22	78.20
LSTM (GloVe)	76.59	75.91	78.37	77.44
CRF (BoW)	71.25	70.40	**73.88**	73.03
CRF (GloVe)	70.23	68.53	72.01	70.48
SVM (BoW)	76.84	76.42	**77.10**	76.68
SVM (GloVe)	68.36	67.94	66.50	65.65

LSTMs, CRFs, and SVMs is explored, and only the highest accuracy rate among a set of hyperparameter configurations is reported per feature set variant in Table 1. Then, we further evaluate the recall, precision, and F1 of the models that achieve the highest predictive accuracy per machine-learning technique. The highest accuracy (79.56%), recall (0.67), precision (0.56), and F1 (0.61) are attained by multimodal LSTMs utilizing the eye gaze features and the bag-of-words method with the post-processing technique applied. Notably, these multimodal LSTMs outperform LSTMs not utilizing eye gaze traces with respect to all the metrics: predictive accuracy, recall, precision, and F1, as well as CRFs and SVMs. A sizable improvement was achieved by the with-gaze LSTMs in the recall rate over without-gaze LSTMs (0.674 vs. 0.642), which indicates multimodal LSTMs are more effective in detecting dialogue breakdowns. This difference accounts for a normalized gain of 8.94%.

4 Conclusion

Conversational pedagogical agents offer great potential for supporting students' problem solving and promoting engagement in game-based learning environments. However, dialogue breakdown between students and agents poses significant challenges and may impede student learning and diminish student engagement. This paper has presented a multimodal deep learning-based dialogue breakdown detection framework that utilizes natural language interactions, eye gaze traces, student gender, and tasks states. Results suggest that a multimodal LSTM-based DBD framework can achieve high predictive accuracies and recall rates, outperforming competitive baseline approaches. In future work it will be important to investigate the potential contribution of additional modalities for improving dialogue breakdown detection. For example, incorporating facial expression and other affective channels may lead to further improvements in dialogue breakdown detection, thereby increasing conversational pedagogical agents' capabilities to engage in even more effective dialogues with students during learning interactions.

Acknowledgments. This research was funded by the National Science Foundation under grants CHS-1409639 and DRL-1640141. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Hirschberg, J., Manning, C.D.: Advances in natural language processing. Science 349, 261–266 (2015).
2. Kim, Y., Baylor, A.L.: Research-based design of pedagogical agent roles: a review, progress, and recommendations. International Journal of Artificial Intelligence in Education 26(1), 160–169 (2016).
3. Tegos, S., Demetriadis, S.: Conversational agents improve peer learning through building on prior knowledge. Educational Technology and Society 20, 99–111 (2017).
4. Graesser, A.C.: Conversations with AutoTutor help students learn. International Journal of Artificial Intelligence in Education 26, 124–132 (2016).
5. Litman, D., Young, S., Gales, M., Knill, K., Ottewell, K., Van Dalen, R., Vandyke, D.: Towards using conversations with spoken dialogue systems in the automated assessment of non-native speakers of English. In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 270–275. (2016).
6. Rus, V., Mello, S.D., Hu, X., Graesser, A.C.: Recent advances in conversational intelligent tutoring systems. AI Magazine 34(3), 42–54 (2013).
7. Lester, J., Ha, E., Lee, S., Mott, B., Rowe, J., Sabourin, J.: Serious games get smart: intelligent game-based learning environments. AI Magazine 34(4), 31–45 (2013).
8. Johnson, W.L., Lester, J.C.: Face-to-Face Interaction with Pedagogical Agents, Twenty Years Later. International Journal of Artificial Intelligence in Education 26(1), 25–36 (2016).
9. Pezzullo, L.G., Wiggins, J.B., Frankosky, M.H., Min, W., Boyer, K.E., Mott, B.W., Wiebe, E.N., Lester, J.C.: "Thanks Alisha, Keep in Touch": Gender Effects and Engagement with Virtual Learning Companions. In: International Conference on Artificial Intelligence in Education, pp. 299–310. Springer (2017).
10. Martinovsky, B., Traum, D.: The error is the clue: breakdown in human-machine interaction. In: Proceedings of the ISCA Workshop on Error Handling in Spoken Dialogue Systems, pp. 11–17. (2003).
11. Higashinaka, R., Funakoshi, K., Inaba, M., Tsunomori, Y., Takahashi, T., Kaji, N.: Overview of Dialogue Breakdown Detection Challenge 3. In: Proceedings of Dialog System Technology Challenge 6 (2017).
12. Higashinaka, R., Funakoshi, K., Araki, M.: Towards taxonomy of errors in chat-oriented dialogue systems. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 87–95. (2015).
13. Steichen, B., Carenini, G., Conati, C.: User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In: Proceedings of the 2013 International Conference on Intelligent User Interfaces, pp. 317–328. ACM (2013).
14. D'Mello, S., Olney, A., Williams, C., Hays, P.: Gaze tutor: a gaze-reactive intelligent tutoring system. International Journal of Human-Computer Studies 70, 377–398 (2012).
15. Hutt, S., Mills, C., White, S., Donnelly, P.J., D 'Mello, S.K.: The eyes have it: gaze-based detection of mind wandering during learning with an intelligent tutoring system. In: Proceedings of the 9th International Conference on Educational Data Mining, pp. 86–93. (2016).
16. Min, W., Mott, B., Rowe, J., Taylor, R., Wiebe, E., Boyer, K., Lester, J.: Multimodal Goal Recognition in Open-World Digital Games. In: 13th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, pp. 80–86. (2017).
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9, 1–32 (1997).

18. Rowe, J.P., Shores, L.R., Mott, B.W., Lester, J.C.: Integrating learning, problem solving, and engagement in narrative-centered learning environments. International Journal of Artificial Intelligence in Education 21, 115–133 (2011).
19. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. Van, Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics 26(3), 339–373 (2000).
20. Min, W., Wiggins, J.B., Pezzullo, L.G., Boyer, K.E., Mott, B.W., Frankosky, M.H., Wiebe, E.N., Lester, J.C.: Predicting Dialogue Acts of Virtual Learning Companion Utilizing Student Multimodal Interaction Data. In: Proceedings of the 9th International Conference on Educational Data Mining, pp. 454–459. (2016).
21. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46 (1960).
22. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543. (2014).
23. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics System Demonstrations, pp. 55–60. (2014).
24. Min, W., Mott, B.W.: NCSU_SAS_WOOKHEE : A Deep Contextual Long-Short Term Memory Model for Text Normalization. In: Proceedings of the Workshop for the Normalization of Noisy User Text, pp. 111–119. (2015).