Dialogue-Based Simulation For Cultural Awareness Training

Sodiq Adewole*, Erfaneh Gharavi*, Benjamin Shpringer†, Martin Bolger*, Vaibhav Sharma†, Sung Ming Yang†, and Donald E. Brown*†

* Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA
† School of Data Science, University of Virginia, Charlottesville, VA, USA

{soa2wg, eg8qe, bs2ux, meb2fv, vs3br, sy8pa, brown}@virginia.edu

Abstract—Existing simulations designed for cultural and interpersonal skill training rely on pre-defined responses with a menu option selection interface. Using a multiple-choice interface and restricting trainees’ responses may limit the trainees’ ability to apply the lessons in real life situations. These systems, also rely on a simplistic evaluation model, where trainees’ selected options are marked as either correct or incorrect. The model cannot capture sufficient information that could drive an adaptive feedback mechanism to improve trainees’ cultural awareness. This paper describes the design of a dialogue-based simulation for cultural awareness training. The simulation, built around a disaster management scenario involving a joint coalition between the US and the Chinese armies. Trainees were able to engage in realistic dialogue with the Chinese agent. Their responses, at different points, get evaluated by different multi-label classification models. Based on training on our dataset, the models score the trainees’ responses for their awareness of the Chinese culture. Trainees also get feedback that informs the cultural appropriateness of their responses. The result of this work showed the following: i) A feature-based evaluation model improves the design, modeling and computation of dialogue-based training simulation systems; ii) Output from current automatic speech recognition (ASR) systems gave comparable end results compared with the output from manual transcription; iii) A multi-label classification model trained as a cultural expert gave results which were comparable with scores assigned by human annotators.

Index Terms—Dialogue-based simulation, cultural-awareness training, intelligent tutoring system, ill-defined domain training, conversational agent, culturally-aware expert system, cross-cultural competence, adaptive feedback system.

I. INTRODUCTION

Dialogue-based interactive simulation is a class of intelligent tutoring system (ITS) often designed for teaching awareness in domains that are not well structured. Examples include: interpersonal skills training [1], cross-cultural competence [2]–[4], negotiation skills [5], [6], confrontation management [7], foreign language learning [8]–[10], and interviewing skills [11].

This paper describes the design and development of an intelligent tutoring system for cross-cultural training. Training to improve cross-cultural interactions is difficult but simulations and games designed for this task have proven to be more effective than traditional classroom instruction [12]. However, most of these systems have used multiple-choice selection or menu-based interface to capture learners’ actions [5], [13]. Menu-based interfaces can be very restrictive and do not reflect the natural mode of human-to-human communication. Multiple choice interaction may also impact the educational efficacy of the training systems [14], [15] by limiting the possibility of the learner being able to transfer any knowledge gained to a real world situation.

Also, the assessment models in existing ITS use simplistic uni-dimensional categorization [2]. These models categorize learners’ actions into one of few categories such as correct/incorrect or negative/neutral/positive. Such simplistic evaluations do not capture the multiple features and dimensions that characterizes interactions in a typical unstructured domains.

This paper expands upon existing ITS by incorporating dialogue-based interaction for cultural competency training. To build this new system required completion of the following tasks:

1) Abstraction of cultural concepts into a computationally tractable feature space;
2) Training multiple expert models to evaluate and score users’ responses in the simulation;
3) Integrating a speech recognition system as the input interface;
4) Implementing an expert assessment model and an adaptive feedback mechanism that work together to provide feedback based on the trainee input; and
5) Designing and conducting an experiment to evaluate the usability and performance of the system.

Since the goal of the our training simulation is to improve the cultural competency of soldiers working with their counterparts from the Chinese army, the ITS described here is built on a Disaster Management Exercise (DME) scenario involving a joint coalition of the US and Chinese army.

II. RELATED WORK

There is a significant amount of literature related to the main elements of this project. By way of organization we have grouped this research into four major areas: Architecture; Interface; Data collection; and User response evaluation.

A. Architecture

A common architecture of intelligent instructional systems consists of four main components. The domain expert model; learner model; pedagogical model; and the user interface. The domain expert model contains the concepts, rules and problem solving strategies for the domain to be learned. It provides
the standard for evaluating learner’s responses. Expert knowledge can be represented in various ways, including network presentations [16], behavior trees [17], [18], and finite state machines [19]. Most instructional systems perform a range of actions such as providing instruction, hint and feedback to the learner [20]–[22] during interaction. The expert model is used to analyze user’s input in order to provide feedback on errors. The learner’s model consists of dynamic representation of the evolving knowledge of the learner. This include aspects (variables) of the learner’s behavior and knowledge that have assumed effect on performance and learning. The Pedagogical model handles the planning and regulation of teaching activities. This includes making decisions on the sequence of activities and strategy to achieve the learning objectives. The plans of this module is based on observations received from the learner about his/her progress as defined in the domain expert model. The user interface is the connecting point between the tutor and the learner. It serves to collect the user’s input as well as render the actions of the tutor. It may come in different forms and media, including a web/mobile app interface or virtual-reality environments.

B. Speech-based dialogue interface

Speech driven interface in human-computer tutoring systems have been shown to help improve learning [14], [15] and also facilitates adaptation of instructional strategy based on perceived affect in learners’ responses [23]. While human tutors can respond to the content of the learners’ response and perceive their emotions [24], existing menu-option selection interface in dialogue-based cultural training simulations limits this capability. The need to close this performance gap, as it applies to simulation systems to learn cultural awareness, coupled with the recent advances in natural language processing, automatic speech recognition (ASR) and text-to-speech systems, drove the investigations presented in this paper.

C. Data collection

Data-driven dialogue systems have mostly relied on Wizard-of-Oz (WOz) [25] or role play simulation [26] methods to gather data. The collected data is then used to train a natural language understanding model to recognize users’ input to the system. Other works have approached data collection using crowd-sourcing through Amazon M-Turk [27]. Sharma V. et al. [27], took a step further by applying a data augmentation technique to improve the sample size and class distribution of an originally crowd-sourced data.

D. Response evaluation and feedback mechanism

Evaluating the trainee’s response, in an intelligent tutoring system, is the main task of the expert model. However, semantic representation and natural language understanding steps needs to be performed prior to this evaluation. Continuous bag of words, term-frequency inverse document frequency (TFIDF) or word vector representation are some of the techniques that have been proposed in the natural language processing literature [28]–[31].

Menu-based option selection systems map each predefined trainee’s response to a corresponding performance measure and feedback [13], [32], [33]. These systems evaluate trainees’ input using supervised machine learning classification algorithm that learns from training examples annotated by subject matter experts [1], [2], [4]. The trained models classifies new input as positive, neutral, or negative; Others use simpler binary category of positive or negative [2]. Applying this annotation method to the case of cultural training simulation means we will define a score as $1 = $Culturally inappropriate$; \ 2 = $Neither appropriate nor inappropriate$; \ and $3 = $Culturally appropriate$. This approach has a number of limitations for our system; i) the complexity of culture-sensitive interaction requires context to properly classify a dialogue. A single score is uninformative and also does not capture any context nor dialogue history; ii) Identifying the difference between the labels may also be a problem as an utterance could fall into more than one category [2]. This happens if a response contains tokens that qualifies it as an appropriate response but also contains an addendum that tends to change the meaning when the utterance is considered as a whole. Humans may still be able to separate such response, the model, however, will be uncertain as to which score to assign such utterance; iv) Adaptive feedbacks are constructed to help learners get better as they interact with the system. Any handcrafted feedback may be off point and not generalize to all possible expressions that the players could make. This may impact the user experience and not provide any information on where the trainee actually erred. This also makes constructing an adaptive feedback almost impossible; v) Lastly, a single score assignment model is not natural of the way human adjudge an utterance for cultural appropriateness. Even though there may be individual differences on interpretation, it is natural to decompose a sentence into components that fits different cultural features. In this paper, we extend this single annotation mechanism into a multi-label model where each label defines a feature that characterizes the utterance based on expectation of the target culture. This method facilitates the decomposition of possible user’s responses into feature space that simplifies the natural language understanding task for the machine learning expert model. Similar to the information state update formulation in task oriented dialogue systems [34]–[37], this method also defines a structure where cultural training dialogues becomes easier to implement in addition to facilitating adaptive feedback. Adapting instruction, hint and feedback [20] to a trainee’s performance in a simulation environment may lead to improved awareness of the target domain [13], [38] and also allows them to know how to subsequently provide more appropriate response [22].

III. Methodology

A. Data collection, annotation and scoring methods

In an earlier work [27] we described the foundation for our data collection effort. Based on the dialogue designed by
the Chinese culture experts [32], players’ in the simulation are evaluated at fourteen (14) different points during the interaction. In order to train the expert models for this task, we crowd-sourced data using Amazon mechanical turk (m-turk). Three approaches were used to prompt m-turk users; i) using context from the scenario, they were asked to provide their own responses to the avatar comments; ii) using feedback provided by the Chinese culture experts in our multiple-choice version [32], they were asked to construct alternative responses; and iii) rephrasing the multiple-choice responses in the earlier version [32] of the simulation. The collected data was annotated by a minimum of two (2) annotators. The annotators had an average inter-rater reliability of 66% [27]. The next section will detail how the annotated data was used in training the models to evaluate and score trainees’ responses in the simulation. The scores assigned by the model are then mapped to the abstracted concepts. This mapping is used to construct a feedback that informed the players of their errors.

1) Cultural Concept abstraction: In [32] subject matter experts provided feedback for each pre-defined option in the system. Using this feedback we developed binary features by abstracting the concepts that the Chinese culture experts used to determine the appropriateness of a response. These features allowed us to cast the problem as a multi-label machine learning classification problem.

Multi-label annotation assigns a score vector with each component being mutually exclusive as follows;

\[
\text{score} = [x_1, ..., x_k] : x_i \in \{0,1\}
\]

This approach makes more sense for a cultural training simulation because; i) In human-to-human interaction, an inappropriate utterance may not be completely inappropriate as there might be elements of such utterance that may still match the interlocutor’s expectation. By decomposing the interlocutor’s expectation into a set of features, the model will be able to separate an utterance based on the natural language representation as discussed in section II-D, and decide whether it is a hit or a miss. ii) Since context is also very important, the features can be defined to implicitly capture what makes an appropriate response within a particular context.

Table I shows example response in one of the sections in the simulation where the player tries to greet the Chinese officer. The features in this section are as follows;

A: Greets officer by saying: "Hi", "Hello" or "Good morning";
B: Avoids asking about officer’s welfare on first meeting; and
C: Uses an honorific expression

This feature-based utterance label allows us to automate feedback construction such that each response is mapped to the features and a feedback could be constructed based on score received from the expert model.

### Table I: Cultural concept abstraction

| Example response                                           | A | B | C |
|------------------------------------------------------------|---|---|---|
| Good morning captain Wang, how are you doing?              |   |   | 0 |
| Good morning captain Wang, how are you doing today?        |   |   | 1 |
| Hello captain Wang, it’s an honor to meet you today.       | 1 | 1 | 1 |

**B. System Architecture**

Figure 1 shows the architecture of the simulation. The modular architecture allowed for easy integration, troubleshooting and error correction. We will describe the interface and each of the components in more detail below:

![Simulation system architecture](image)

1) User Interface: Virtual environment: The simulation environment was designed using the Unity game engine. Characters were modeled and rigged using Adobe fuse while animation was done using Mixamo 3D character animator. Lip-syncing and facial animation helped bring the characters to life and created a more realistic interaction in the virtual environment. Figure 2 shows one of the American soldier character in the virtual environment.

![Virtual simulation environment](image)

Voice Detection (VDM) and Automatic Speech Recognition (ASR) Modules: The voice detection module is a gated listening mechanism that manages turn taking between the player and the avatar. It uses a sampling technique [39] to
to repeat what (s)he has said again.

If the ASR module confidence level is below this threshold, the avatar will ask the player to repeat him/herself.

The ASR module is configured to output both the transcribed text and the confidence parameter $\alpha$ for the recognized speech output. First, we tried different values of $\alpha$ to balance the trade-off between frustrating users and the accuracy of the expert model. This allows the speech recognizer to only pass output above $\alpha$ to the expert module. If the ASR module confidence level is below this threshold, the dialogue loops back and the avatar would ask the player to repeat what (s)he has said again.

2) Cultural-Expert Module: This module houses the pre-trained input-vectorizer and the multi-label classifier models as shown in figure 3. We trained multiple multi-label classifier models as our domain expert to evaluate and score the player’s response for each section. We considered the problem of recognizing whether a user’s utterance is culturally appropriate or not as a multi-label text classification problem [40], [41] as described in section II-D. Each user utterance is scored based on the defined features for the section. Since the goal of the simulation is to improve cultural awareness of soldiers working with their counterparts from the Chinese army, we built the scenario around a Disaster Management Exercise (DME) involving a joint coalition effort between the US and the Chinese army. Chinese culture experts designed a dialogue tree structure with fourteen (14) strategic points where the players’ responses are evaluated for cultural awareness. For improved accuracy, each of these evaluation tasks are handled by separate evaluation models and trained on separate dataset as described in section II-C.

**Classifier models:** The models were trained on our annotated dataset of data as detailed in [27]. We experimented with different classifier models as well as different combinations of the input vector dimensions to determine the best performing model. We evaluated the performance of the models based on the F1-score, precision and recall. We obtained the best performance with k-nearest neighbours (KNN), RandomForest and multi-layer perceptron (MLP) classifier models. RandomForest and multi-layer perceptron (MLP) models were integrated into the final simulation because, while KNN had a good performance, the computational cost during the simulation was too expensive and the time taken for the player’s response to be evaluated makes the deployment less suitable for a real-time interactive simulation. Our evaluation metric was carefully chosen for reasons that captures both pedagogical objective as well as the user-experience considerations. Below, we detail suitability of each of these metric:

- **Precision:** is calculated as the ratio of the true positive (TP) to the sum of the true positive (TP) and false positive (FP). High precision indicates high TP and low FP which means that the expert-module mostly recognizes when the users make the appropriate. Also, low FP means the players were not wrongly assigned a score when they gave an inappropriate response. This is bad for pedagogical efficacy of the system since the users will not be corrected for what they did wrongly. Minimizing the FP is important for the pedagogical efficacy of our system.

- **Recall:** is the ratio of the TP to sum of TP and false negatives (FN). High recall indicates high TP to FN ratio which is desirable for a good user experience in the simulation. This means that the classifier mostly recognizes when the users gives the appropriate response and scores them properly for it. Since the feedback received is based on what the output score is, low false negative means the users are not frustrated by the system correcting them even when they make an appropriate response. We wanted a good recall rate so players will not receive feedback that says their response is not appropriate, even when it is. Inaccurate feedback is bad for
educational effectiveness of the interaction as players may get confused as to what actually defines an appropriate response. **F1 score:** is the harmonic mean of precision and recall metrics. The F1 score is an important metric in this system as it aggravates the lower of the outlying value of the precision and recall metrics. It offers information as to the trade-off we are making between pedagogical efficacy and the user experience of the simulation. Precision and recall can individually be maximized by minimizing false positives and false negatives respectively.

**ASR Word Error Rate:** measures the performance of the speech recognition module. This metric will help us understand and compare the performance of the expert models. Since the model performance is dependent on whether or not it got the correct token from the ASR system. If the expert module is not finding the token that identifies a user response as appropriate due to error from the speech recognizer, the expert module will consider this response as inappropriate and may assign a score of zero for that feature.

3) **Pedagogical module::** Two pedagogical components were implemented to help players navigate interactions with the avatars. A simulation guide help players navigate the interaction by providing background information about the non-player characters. The feedback module informs the players of what was right and wrong about their response. We will describe both components in detail below:

**Simulation guide:** The simulation guide is a pre-recorded audio with corresponding text display as shown in both figures 2 and 4. This provides the players with context of the scenarios and help guide them by providing additional information that the players need to properly construct their responses. This is important for several reasons; i) it helps them clarify and contextualize what was said by the avatar and also connects it to the history of the conversation. This will help the players recall the history of the dialogue in the previous scenes and connect to the current one; ii) it helps narrow the space of potential response so the players could articulate their response within a shorter time; iii) it also helps the flow of the interaction when the player are able to respond promptly; iv) It increases the players immersion in the scenes and also their likelihood of making the proper response; and v) For non-military players who may not be familiar with the scenario, the guide helps them understand the context of what they needed to deal with.

**Adaptive feedback generation** Adaptive feedbacks based on intelligent error analyses of learners’ solutions can help achieve correct response with minimal error as learners progress through the simulation [22], [38]. This is especially useful for our simulation because, it help players continuously improve their responses as they try to implement the corrections in subsequent turns [43]. Our adaptive feedback mechanism uses the output score-vector as shown in equation 1 from the expert module and maps it to the domain feature space of the dialogue. Depending on the score output, the feedback generator constructs a feedback that informs the player of what an appropriate response should be and informs them of what they did right and tries to correct what they could improve upon in subsequent dialogue.

An example feedback for the first response from table I which is an example utterance that a player could say when meeting the Chinese captain for the first time; (i.e.“Good morning captain Wang, how are you doing?”). The domain-expert module returns a score vector of [1,0,0] based on the following features;

(i) “Greets the officer” - [1]
(ii) “Avoids asking about officer’s welfare” - [0]
(iii) “Uses an honorific expression” - [0]

the feedback generated is as follows:

“a culturally appropriate response in this section should include greeting the officer, avoiding asking about the officer’s welfare on a first meeting, and using an honorific expression. From your response, you succeeded in greeting the officer, but your response could be improved by avoiding asking about the officer’s welfare on a first meeting and using an honorific expression.”

The generated feedback text is passed to the feedback speech synthesizer to get converted to audio format and played to the players before proceeding to the next turn.

IV. EXPERIMENT

We recruited 18 subjects to test the usability and performance of the simulation. We evaluated the performance of the models and the overall system as a potential training tool for cultural awareness. Participants were mostly undergraduates from the school of engineering at the University of Virginia. The objective of our simulation system design was to explore culturally-aware data-driven model that evaluates the speech-based dialogue between the user and the computer avatar. Our experiment did not cover testing learning improvement in the target culture as this was outside the scope of this paper. The experiment was approved by the IRB and all participants provided their consent for the data that was collected.

Our hypotheses are as follows:

**H1:** The expert model will not assign scores comparable with
human adjujged scores based on the players’ responses in the simulation.

**H2:** The word error rate of the speech recognizer does not have any impact on the performance of the expert model when compared with manually transcribe input.

**H3:** The adaptive feedback mechanism aids the user in making more culturally appropriate responses as they progressed through the interaction.

**H4:** An integrated interactive simulation system with speech-driven input designed for cultural awareness training created room for more naturalistic interaction with better user experience.

To analyse the performance of the cultural-expert models, we collected the text transcript of the participants’ responses as well as the audio recordings. This logged text also include the score assigned by the expert module and the ASR confidence level. For the user experience, we administered questionnaires to the participants to understand what went right and how the system can be improved.

V. RESULTS

The first five subjects’ data were removed because the system did not use the final versions of our classification models. This leaves the data for 13 participants. The system froze twice and the data collected for the two (2) participants was omitted in the analysis as well, leaving data for eleven (11) participants.

### A. Expert Model performance

The recorded audio responses were manually transcribed, the result was then annotated by two human annotators. The annotators had inter-rater agreement of 78% and the score assigned was compared with the output scores from the expert models. Using the score from the human annotators as a base-line to calculate the F1-score, recall, and precision. We also considered the ASR system word error rate to explain in part what was observed in the performance of the models.

| Model Id | No of features | F1 (%) | Precision (%) | Recall (%) | ASR WER (%) |
|----------|----------------|--------|---------------|------------|-------------|
| 1        | 3              | 94.1   | 88.9          | 100.0      | 63.2        |
| 2        | 3              | 80.0   | 69.2          | 94.7       | 25.4        |
| 3        | 2              | 93.3   | 87.0          | 100.0      | 17.3        |
| 4        | 3              | 75.0   | 68.2          | 83.3       | 19.7        |
| 5        | 2              | 79.1   | 77.3          | 81.0       | 15.0        |
| 6        | 3              | 86.7   | 96.3          | 78.8       | 17.3        |
| 7        | 3              | 63.4   | 65.0          | 61.9       | 14.6        |
| 8        | 4              | 52.9   | 45.0          | 64.3       | 15.3        |
| 9        | 3              | 80.0   | 72.0          | 90.0       | 8.6         |
| 10       | 2              | 88.0   | 100.0         | 78.6       | 19.3        |
| 11       | 1              | 88.9   | 80.0          | 100.0      | 4.4         |
| 12       | 2              | 83.3   | 100.0         | 71.4       | 19.9        |
| 13       | 3              | 89.4   | 95.5          | 84.0       | 19.2        |
| 14       | 2              | 76.9   | 66.7          | 90.9       | 16.4        |

**TABLE II** Performance of cultural-expert module

Figure 5 shows the plot of the model performance in table II. The expert models had average F1-score of 80.7%, average precision of 79.4% and recall of 84.4%. The average WER of the ASR system is 19.6%.

**H1:** The F1-measure of the expert models as shown in table II shows average performance of 80%. Except for model 7 and 8 with 3 and 4 features respectively, other models had F1-score above 75%. Models 7 and 8 had more binary features resulting in \(2^k\) classes into which every response could fall into. From this we see that the number of features impacts the performance of the models as the classification task gets more difficult with increasing number of features.

The high WER for section 1 is because participants were mostly mentioning the name of the Chinese avatar (“Captain Wang”) with whom they were speaking and the ASR system struggled to properly recognize it. Training the ASR system on our own dataset may have improved the performance in this section. With average precision of 79.4% the ratio of true positives (TP) to false positives is well above average which shows that the model does not assign score when it receives an inappropriate response from the players. Similarly, average recall of 84.4% showed that the models TP to false negatives (FN) is around 0.84. This also indicates positive user experience performance as the players do not get corrected for what should pass as a culturally appropriate response. Both measures shows high TP values which is good for the pedagogical efficacy of the simulation. Since low recall values indicates higher FN, models 7 and 8 with the least recall means that players were, on average, corrected more than a human evaluator would have corrected similar utterance. Based on F1-score \((t(13) = 26.49, p=1.07e-12)\), we have enough evidence to conclude that the performance of the expert module was comparable with that of human expert using manually transcribed data.

**H2:** We performed simple regression analysis to analyse the...
relationship between the WER the F1-score. Our result \( t(13) = 1.146, p = 0.274 \) showed no significant \((p>0.05)\) relationship exist between the word error rate of the speech recognizer and the performance of the expert models. Similarly, the performance of the ASR system had no impact on the other metrics (Precision and Recall) of the expert models. This shows that, the tokens needed to score the participants’ responses were mostly properly recognized. For example, model-1 where participants get score when they greeted the army officer that they were interacting, tokens such as “Good morning”, “hi”, ”hello” were mostly correctly recognized and appropriately scored by the model. However, a multiple regression analysis using the number of features together with the WER to predict the F1-score of the models showed that both variables have joint significant effect \( F(2,11) = 6.179, p=0.016 \) on the performance of the models. Therefore, as the number of features increases, the effect of the speech recognizer error tend to begins to have effect the model performance.

**H3:** Using average aggregate scores across sections (equation (2)), we evaluated the performance of the participants as they progress through the simulation. We hypothesized that providing an adaptive feedback to the players based on the model assigned score will improve their performance as they progress through the interaction. Natural human-to-human interaction does not usually provide explicit feedback to help players identify errors in their responses. However, it was still necessary to provide feedback to participant to help them learn of their mistakes based on the norm in the target culture.

\[
Score_{avg} = \frac{1}{N} \sum_{j}^{N} \left( \frac{1}{k} \sum_{i}^{k} x_i \right)
\]  

(2)

From equation (2), \( N \) is the number of participants and \( k \) is the number of cultural concept that each model was evaluating in the users’ utterance; \( k \) is defined for each model as the number of labels/features as indicated in table II.

![Fig. 6: Participants average score](image)

From Figure 6 above, we did not observe any performance improvement in the aggregate score received by the participants during the interaction. This is may be due to the independence of the features at each evaluation points. Also, most of the participants have not had any prior experience using such simulation system for cultural training and this may impact their performance as a first timer. The participants may also find it difficult remembering what corrections they received in previously sections. Also, lessons learnt in one simulation run can be easily utilized experientially in a subsequent one but not necessarily during the same interaction. Players would have had time to internalize the corrections from the first experience and if they have to go through the simulation on a second run, their experience may be better.

**VI. CONCLUSION**

In this paper we presented a data-driven simulation system for cultural awareness training. We improved previous simulation systems by implementing a spoken dialogue system in place of the menu-based option selection model. We implemented a data-driven multi-model cultural experts to
evaluate and assign score to users’ spoken responses based on abstracted cultural features. Our design also implemented an adaptive feedback mechanism that uses the score assigned by the expert module to serve the pedagogical role of informing the players how culturally appropriate their responses were. We evaluated the system by comparing the performance of the models with scores assigned by two human annotators. Our result showed that our data-driven intelligent experts models gave comparable performance with what human would adjudge similar utterance. The feature abstraction technique and adaptive feedback mechanisms also allowed us to solve the problem of manually scoring user’s utterance. We also did not have to write feedback for every possible utterance that a user could make in the simulation. These components also allowed for real time adaptive feedback construction based on vector-score received by the learners’ utterance. Most prior works have focused on developing techniques to improve the individual component of the intelligent tutoring system, our simulation focused on improving most of the modules to improve the realism of the interaction. While, providing explicit feedback in an interactive situation does not reflect the natural way in which learning occurs during interaction, our future works will explore techniques towards an adaptive response generation with implicit feedback. This will have the ultimate goal of improving the learning efficacy of cultural training simulation.

ACKNOWLEDGEMENTS

The project is funded by the Army Research Lab under Research Funding Source Award Number: W911NF1820279. We thank the Data Science Institute of the University of Virginia for the support received for this project.

REFERENCES

[1] H. C. Lane, M. G. Core, D. Gomboc, A. Karnavat, and M. Rosenberg, “Intelligent tutoring for interpersonal and intercultural skills,” UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE …. Tech. Rep., 2007.
[2] H. C. Lane and M. J. Hays, “Getting down to business: Teaching cross-cultural social interaction skills in a serious game,” in Workshop on Culturally Aware Tutoring Systems (CATS), 2008, pp. 35–46.
[3] P. Dillenbourg, D. Schneider, and P. Synteta, “Virtual learning environments,” in 3rd Hellenic Conference Information & Communication Technologies in Education”. Kastaniotis Editions, Greece, 2002, pp. 3–18.
[4] H. C. Lane, M. J. Hays, M. G. Core, and D. Auerbach, “Learning intercultural communication skills with virtual humans: Feedback and fidelity.” Journal of Educational Psychology, vol. 105, no. 4, p. 1026, 2013.
[5] J. M. Kim, R. W. Hill Jr, P. J. Durlach, H. C. Lane, E. Forbell, M. Core, S. Marsella, D. Pynadath, and J. Hart, “Bilat: A game-based environment for practicing negotiation in a cultural context,” International Journal of Artificial Intelligence in Education, vol. 19, no. 3, pp. 289–308, 2009.
[6] M. Core, D. Traum, H. C. Lane, W. Swartout, J. Gratch, M. Van Lent, and S. Marsella, “Teaching negotiation skills through practice and reflection with virtual humans,” Simulation, vol. 82, no. 11, pp. 685–701, 2006.
[7] J. Kolkmiejer, M. Lee, and D. Heylen, “Moral conflicts in vr: Addressing grade disputes with a virtual trainer,” in International Conference on Intelligent Virtual Agents. Springer, 2017, pp. 231–234.
[8] M. L. Swartz and M. Yazdani, Intelligent tutoring systems for foreign language learning: The bridge to international communication. Springer Science & Business Media, 2012, vol. 80.
[9] H. Wang, C. J. Waple, and T. Kawahara, “Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition,” Speech Communication, vol. 51, no. 10, pp. 995–1005, 2009.
[10] P. Wik and A. Hjalmarssson, “Embodied conversational agents in computer assisted language learning,” Speech communication, vol. 51, no. 10, pp. 1024–1037, 2009.
[11] Z. Yu, V. Ramanarayanan, P. Lange, and D. Suendermann-Oeft, “An open-source dialog system with real-time engagement tracking for job interview training applications,” in Advanced Social Interaction with Agents. Springer, 2019, pp. 199–207.
[12] D. E. Brown, A. Moenning, S. Guerlain, B. Turnbull, D. Abel, and C. Meyer, “Design and evaluation of an avatar-based cultural training system,” The Journal of Defense Modeling and Simulation, vol. 16, no. 2, pp. 159–174, 2019.
[13] K. Georgila, M. G. Core, B. D. Nye, S. Karumbaiah, D. Auerbach, and M. Ram, “Using reinforcement learning to optimize the policies of an intelligent tutoring system for interpersonal skills training,” in Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 737–745.
[14] D. J. Litman and S. Silliman, “ITSPOKE: An intelligent tutoring spoken dialogue system,” in Demonstration papers at HLT-NAACL 2004. Association for Computational Linguistics, 2004, pp. 5–8.
[15] R. Moreno, R. E. Mayer, H. A. Spires, and J. C. Lester, “The Case for Social Agency in Computer-Based Teaching: Do Students Learn More Deeply When They Interact With Animated Pedagogical Agents?” Cognition and Instruction, vol. 19, no. 2, pp. 177–213, Jun. 2001. [Online]. Available: https://doi.org/10.1207/S1532690XCI1902_02
[16] H. Gamboa and A. Fred, “Designing intelligent tutoring systems: a bayesian approach,” Enterprise Information Systems III. Edited by J. Filipe, B. Sharp, and P. Miranda. Springer Verlag: New York, pp. 146–152, 2002.
[17] X. Geng, S. Qin, H. Chang, and Y. Yang, “A hybrid knowledge representation for the domain model of intelligent flight trainer,” in 2011 IEEE International Conference on Cloud Computing and Intelligence Systems. IEEE, 2011, pp. 29–33.
[18] T.-W. Chan, “Curriculum tree: a knowledge-based architecture for intelligent tutoring systems,” in International Conference on Intelligent Tutoring Systems. Springer, 1992, pp. 140–147.
[19] A. Yankovskaya and N. Yevtushenko, “Finite state machine (fsm)-based knowledge representation in a computer tutoring system,” New
Below is the transcript of the dialogue of one of the participants for the first three scenes. We showed the score received (in parenthesis)

**Scene 1:**

**Captain Heist:** Good Morning. As you know, we are scheduled to meet our Chinese counterparts for the first time today. It is important that you make a good first impression. Unfortunately, I have had a requirement come up last minute and will not able to meet the Chinese captain today. I am having you go in my place. Please receive him cordially and begin making plans with him as soon as possible.

**Participant:** Thanks for the information.

**Scene 2:**

**Participant:** Good morning Captain Wang.

**Captain Wang:** Good morning, Lieutenant.

**Participant:** Are you the leader of the Chinese component of the coalition.

**Captain Wang:** Yes, I am. I made arrangements with Captain Heist. Is he here?

**Participant:** Unfortunately Captain Heist had prior duties and I will be taking his place.

**Captain Wang:** Why couldn’t he come?

**Participant:** He unfortunately had prior engagements with other officers.

**Captain Wang:** I understand. Let’s begin preparing because we need to leave soon.

**Participant:** Great let’s get started sir.

**Scene 3:**

**Participant:** Captain Wang is there anything I can help you with?

**Captain Wang:** I’m Fine, Lieutenant. I just had some questions about the mission.

**Participant:** Well what are your concerns officer.

**Captain Wang:** Well what are your concerns officer.

**Participant:** Great let’s get started sir.

**Scene 4:**

**Participant:** Did Captain Wang is there anything I can help you with?

**Captain Wang:** I’m Fine, Lieutenant. I just had some questions about the mission.

**Participant:** Well what are your concerns officer.

**Captain Wang:** Well what are your concerns officer.

**Participant:** Great let’s get started sir.

**Scene 5:**

**Participant:** Did Captain Wang is there anything I can help you with?

**Captain Wang:** I’m Fine, Lieutenant. I just had some questions about the mission.

**Participant:** Well what are your concerns officer.

**Captain Wang:** Well what are your concerns officer.

**Participant:** Great let’s get started sir.
supply depot where you keep everything? I don’t want to bother you any longer, so if it will save you time, I’ll walk over myself.

Participant: Well Captain the Supply Depot is down the way to the left so if you want to check there that would be a good hint to so it will bring.

Captain Wang: Oh, well thank you Lieutenant! Good luck on the planning.

Participant: I will be around until the mission brief this afternoon if you need me at any point.