Data Collection Methods for Building a Free Response Training Simulation

Vaibhav Sharma, Beni Shpringer, Sung Min Yang, Martin Bolger, Sodiq Adewole, Dr. D. Brown, and Erfaneh Gharavi
University of Virginia, vs3br, bs2ux, sy8pa, meb2fv, soa2wg, brown, eg8qe@virginia.edu

Abstract – **Most past research in the area of serious games for simulation has focused on games with constrained multiple-choice based dialogue systems. Recent advancements in natural language processing research make free-input text classification-based dialogue systems more feasible, but an effective framework for collecting training data for such systems has not yet been developed. This paper presents methods for collecting and generating data for training a free-input classification-based system. Various data crowdsourcing prompt types are presented. A binary category system, which increases the fidelity of the labeling to make free-input classification more effective, is presented. Finally, a data generation algorithm based on the binary data labeling system is presented. Future work will use the data crowdsourcing and generation methods presented here to implement a free-input dialogue system in a virtual reality (VR) simulation designed for cultural competency training.**

Index Terms – Cultural Competency Training, Data Crowdsourcing, Data Generation, Virtual reality simulation

INTRODUCTION

Past research in the area of cross-cultural competence training has shown that simulation-based training systems, in which users interact with an avatar from the target training culture, are more effective than non-interactive class or book-based training [1]. Despite the demonstrated educational advantages of simulation-based training systems, most previous research in this area has focused on simulations with unrealistic multiple-choice-based dialogue systems [2][3]. A free-response-based training system would allow the user to realistically interact with the avatar, potentially increasing the training effectiveness.

The overall goal of this research is to implement and test the effectiveness of a simulation for training US military officers to productively interact with their Chinese counterparts. On-screen multiple-choice options in an existing cross-cultural competence training simulation will be replaced with free response input collected using a speech-to-text system and a text classifier.

Investigating an effective method of cross-cultural competence training for Chinese-American interactions is a valuable research aim for two key reasons:

First, there are major cultural differences between the US and China. One way of measuring cultural differences is psychologist Geert Hofstede's cultural dimension theory [4]. Hofstede measures cultural differences along six dimensions: Power Distance, Individualism, Uncertainty Avoidance, Masculinity, Long Term Orientation, and Indulgence vs. Restraint [5]. The US and China display large differences in the Power Distance and Individualism dimensions, which makes cross-cultural competence training more likely to be valuable [6].

Second, US intercultural interaction with Chinese civilian and military populations continues to increase according to several metrics. For example, the Chinese immigrant population in the US has grown more than six-fold since 1980 [7]. The US military has participated in the annual joint Disaster Management Exchange (DME) with the Chinese military since 2005 [8]. The goal of DME is to improve the ability of the US military to respond to natural disasters in the Pacific region by training cooperatively with the Chinese military.

A key challenge when creating a free-input training simulation is obtaining sufficient data to train classification models that will have the ability to replace the multiple-choice dialogue system. In this paper, methods for collecting, generating, and labeling data for creating a free-input dialogue system are presented.

These methods have the ability to produce a robust corpus for accurately classifying speech in a training simulation. While a great deal of effort is required to build such a corpus, this framework could be effective for a variety of free-response classification problems, especially in the realm of avatar-based behavioral training simulations.

LITERATURE REVIEW

I. Frameworks for Cultural Dimensions and Politeness

Besides Hofstede's dimension theory, other metrics such as politeness that could act as a proxy for cross-cultural competence (C3) have been extensively investigated. Past work on generalized models for politeness include Brown and Levinson's politeness theory [9]. This theory focuses on the concept of saving 'face' (i.e. social approval and acceptance) as a motivating force for individuals in social interactions [9][10]. According to Gu, face in Chinese culture centers on respecting social norms rather than psychological desires [10][11]. Politeness frameworks have been empirically tested [11] and used to create realistic dialogue simulations with the ability to judge abstract attributes like familiarity and solidarity [12].

II. The Case for Immersive Simulations

According to Selmeski, the common forms of training for 3C in militaries include pre-deployment briefings, awareness training, and pocket references [13]. Lane argues that such approaches rely significantly on rote learning with little effectiveness in teaching an understanding of another culture [14]. He argues that cultural competence comes from a "heightened sense of self-awareness, an ability to self-assess, enhanced perceptive abilities, and a proclivity to reflect on experience", which he summarizes as metacognitive maturity [14]. Lane proposes that immersive learning environments are a more effective way to train for 3C and finds promise in virtual reality environments combined with machine learning as a potential avenue for 3C training [14].

III. Immersive Educational Environments

In line with Lane's work, there has been increased research into using immersive environments for training 3C. Previous attempts to train C3 using virtual reality have shown encouraging results. For example, Roberts et al. used a simulation environment using actors to train for 3C in a nursing context [15]. In 2014, the Cultural Awareness in Military Settings Project created a simulation of an Afghanistan village to train Norwegian personnel [16]. In 2018, Sheridan et al. produced a VR simulation emulating a DME [17]. The game's dialogue was produced under the guidance of Chinese language and culture experts [17]. Within the game, one goes through the exercise by selecting multiple-choice dialogue options. Users receive feedback depending on the appropriateness of the option selected [17].

IV. Data Collection for Dialogue Systems

Data set collection is an important aspect of the creation of dialogue classification systems [18]. Kang et. al. conducted comparisons of various methods for data collection using crowdsourcing [19]. Each method is evaluated based on its ability to generate a data set that covers the feature space of the problem, provide a diverse sample, and perform well on a classification task. Past research has focused on the extraction of text features from large scale online corpuses. These text features are then used to train a variety of models to produce general conversational chatbot systems [20]. The use of crowdsourcing of labels for text has also been previously investigated [21]. This work builds on previous work by introducing a novel labeling scheme and an accompanying data generation approach. The proposed data generation approach and labeling scheme can be used in conjunction with data set crowdsourcing.

APPROACH

I. Introduction to Test Data Crowdsourcing

Table 1 gives an example of a multiple-choice options from the original game. The user is asked about the type of supplies they are planning to bring on a mission.

TABLE I
EXAMPLE MULTIPLE CHOICE RESPONSES AND FEEDBACK

Multiple Choice Responses	Feedback
I appreciate you looking out for your team, but you will get all of it soon when we have the brief this afternoon.	Moderately appropriate, displays understanding and empathy towards the officer and his need to know, but is direct in telling him he will not comply.
Is it alright if this waits until the actual mission brief? I'm busy getting ready for it now.	Has an air of politeness to it that implies if the Chinese officer really needed the information, that you will give it, but it would be inconvenient to do so.
Not all of the details have been finalized, so I'm not sure if that it would be of any help to you now.	This response is curt and dismissive.

After the user answers, feedback is given. The educational objective is to teach the user Chinese/American cross-cultural communication stress points. The feedback is the training mechanism.

One issue with converting a multiple-choice based dialogue system to a free-input based dialogue system is that the context provided by the multiple-choice answer options is lost. For example, the user is not told that details for the supplies have not been finalized yet, but they learn that this is true because it is given as an answer option. In the free-input version of the game, this is handled by giving the user a short briefing before they answer. The pre-question brief explains their character's motivation and gives additional information that helps the user formulate a reasonable response.

Given that there are many ways of responding to the prompt, a large data set of responses needs to be collected. To do this, Amazon Mechanical Turk (AMT) was used to crowdsource data.

II. The Three Types of Test Data

Three types of prompts were used to collect data: rewrite prompts, feedback prompts, and context prompts. All three prompt types asked the user to provide a response to the dialogue from the original simulation. Users were given the context necessary to understand parts of the dialogue in isolation.

Rewrite prompts asked the user to rephrase each the original multiple-choice options. Example rephrasings of the second piece of dialogue on Table I include:

- "I understand your concern. Once we have our afternoon brief, you will get all the supplies you requested."

- "Let's wait till the actual mission brief, if that is okay with you? Right now, I am busy getting everything in order."

Feedback prompts asked to user to provide a response that matches the feedback for each original multiple-choice response. For instance, users were asked to write a response that "is curt and dismissive". Example responses for this prompt included:

- "We can't disclose that at this time."
- "I am not at liberty to release that information to you Commander."

The intention was to give users more freedom when answering. This could lead to more lexically diverse responses.

Context prompts only provide users with the pre-question brief from the game. They differ from the feedback and rephrasing prompts because they do not give example answers. These prompts only gave a short explanation of the motivation and context necessary to answer the question. The information provided in context prompts is the same as the information users are provided in the game. The motivation for using context prompts was to collect diverse responses that were likely to be similar to what a simulation user would actually say.

For the question in table I, users were given the brief: "The Chinese captain is interested in coordinating with you and helping your team. However, you are currently busy preparing for a brief with the Chinese team. You will be covering the information he is asking about during the meeting, but you have not finalized all the details yet."
Example responses included:

- "Things are moving quickly on this mission. The final details will be available this afternoon."
- "I'd be happy to go through the information on the types of supplies in our meeting later this afternoon. Right now, unfortunately, I have other matters to attend to."

The examples above generally correspond with the first and second pieces of dialogue in Table I. Thus, users were able to create reasonable responses without the examples the rewrite and feedback prompts provided.

III. Binary Data Labeling

Creating accurate feedback for responses is another challenge faced when converting a multiple-choice dialogue system to a free-input dialogue system. The original feedback categories are too restrictive to cover the potential feature space of responses. To deal with this issue, a binary category labeling system was implemented. Each dialogue option was classified into its own set of binary categories. The binary categories were created to label the presence or absence of independent features in the text. The labels determine the cultural appropriateness of each response. The labels used to classify each question were created by looking at the original feedback for each question and determining the educational objective.

Labeling was completed by judging each response based on the binary categories. For example, Table II shows how the first piece of dialogue on Table I could be scored:

TABLE II
BINARY LABELS FOR EXAMPLE FREE-RESPONSE

Example response	Gives time when information will be given	Displays understanding and empathy	Explains that you are not ready
Is it alright if this waits until the actual mission brief? I'm busy getting ready for it now.	1	0	1
I appreciate you looking out for your team, but you will get all of it soon when we have the brief this afternoon.	1	1	0

IV. Label Validation

Each response was labeled by two raters. Inter-rater reliability was calculated using Cohen's Kappa. The average kappa for all questions in our data set is 0.613 with variance of 0.033. Viera et al. report that kappa values of 0.41-0.6 indicate moderate agreement and values of 0.61-0.8 indicate substantial agreement [21]. When raters agreed about the labeling for the response, the response was added to the data set for training a classifier. If the raters did not agree on the label for at least one binary category, one of the rater's labels was randomly selected.

V. Training Data Generation Motivation

Breaking the data labels into independent binary categories makes it possible to generate a large training data set. It is assumed that the binary category labels are applied based on a subset of the text in each input. One approach to generating data would be to write a series of snippets of text that definitely fit into each of the categories. This approach is based on the assumption that the presence of the snippet in an input sentence implies that the input sentence should always receive a specific binary label for that category. This assumption is invalid because the meaning of a snippet of the text could be altered by adjacent text. For example, when asked to explain someone's absence, isolated explanations such as "she is unavailable" and "she was unable to come" receive a binary label of 1 and isolated statements such as "she is not here" or "she is not coming" receive a label of 0. A problem arises when two of these statements are combined. For example, "she is not here because she was unable to come" should receive a 1. If two phrases never appear together in a training data set, a model may not easily be able to predict the correct label when they appear together.

VI. Training Data Generation Categorization

To capture this in the training set, data is generated by creating combinations of text snippets. Let $t_j(i)$ be the i^{th} snippet of example text in category j. It is not assumed that the categories contain the same amount of data. Full input text can be created by concatenating a text snippet from all or a subset of the categories. The text snippets belong to four possible classes: def(k) and iso(k) where $k = 0{,}1$ represents the binary label for the elements in the class. $t_j(i) \in \text{def}(k)$ for $k = 0, 1$ if $t_j(i) \in text$ always implies $text$ will receive binary label k for category j. Similarly, $t_j(i) \in \text{iso}(k)\ for\ k = 0\ or\ 1$ if $t_j(i) + t_j(h) \in \text{def}(k)$ for any $t_j(h) \in \text{def}(k)$. Therefore, text snippets in $iso(k)$ are dominated by text snippets in def(k). Given that def(k) is dominate over $iso(k)$, four additional sets can be created by concatenating $t_j(i) \in iso(k)\ \forall\ i$ with $t_j(h) \in def(k)\ \forall\ h$. This results in four additional sets: $\text{def}(k) + iso(k)\ for\ k = 0{,}1$. Members of these sets satisfy the definition for membership in $def(k)$, so they are subsets of $def(k)$.

VI. Training Data Generation Categorization

def(0):

Binary Cat. 1	Binary Cat. 2
a	d
b	e
c	

def(1):

Binary Cat. 1	Binary Cat. 2
f	h
g	i
	j

Figure I
DATA GENERATION ALGORITHM EXAMPLE TABLES

The elements of each table in Figure I represent text snippets that belong to each corresponding set. The number of rows is equal to the max number of text snippets in a category across both tables. The number of columns is given by the number of categories. Using the notation from above, $t_1(1) = a$.

The algorithm takes a column number and a blank string as input. It starts by appending each element in the first column to a list and recursively calling the function on the next row. It does this with a separate loop for each table. This process can be illustrated by the tree like structure shown in Figure II:

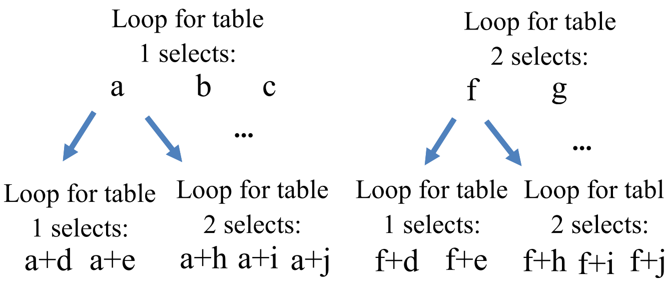
Loop for table 1 selects:
a b c
...
Loop for table 2 selects:
f g
...
Loop for table 1 selects:
a+d a+e
Loop for table 2 selects:
a+h a+i a+j
Loop for table 1 selects:
f+d f+e
Loop for table 2 selects:
f+h f+i f+j

Figure II
DATA GENERATION TREE STRUCTURE

The recursive function outputs when the number of concatenated text snippets equals the number of categories. Categories can be left blank by inserting blank rows into the table.

RESULTS

I. Test Set Visualizations

Comparing the lexical diversity and average response length of each of the types of test data provides a measure of the utility of each data type. Due to the data set size imbalance across the types of data (see the legend of Figure III), Measure of Textual Lexical Diversity (MTLD) was used to measure lexical diversity. Types to tokens ratio (TTR), the standard measure of lexical diversity, is biased toward higher values for smaller corpuses. MTLD differs from TTR because it calculates "the mean length of sequential word strings in a text that maintain a given TTR value" [21]. Koizumi found that MTLD is less affected by text length than other measures of lexical diversity [21].

In Figure II, the horizontal and vertical bars give the mean MTLD and average length over all the questions in the data set. It indicates that context prompts yield longer answers with higher MTLD values. For some questions (e.g. 3 and 4), there are large differences in average response and MTLD value. Each of these questions asks the user to provide and explanation, and the context prompts encouraged users to provide more verbose explanations in an attempt to be more polite. The values for other questions, such as questions 1 and 12, are relatively similar regardless of the prompt type. Both of these questions are greetings, so the clustering could be due to the small feature space that these questions have.

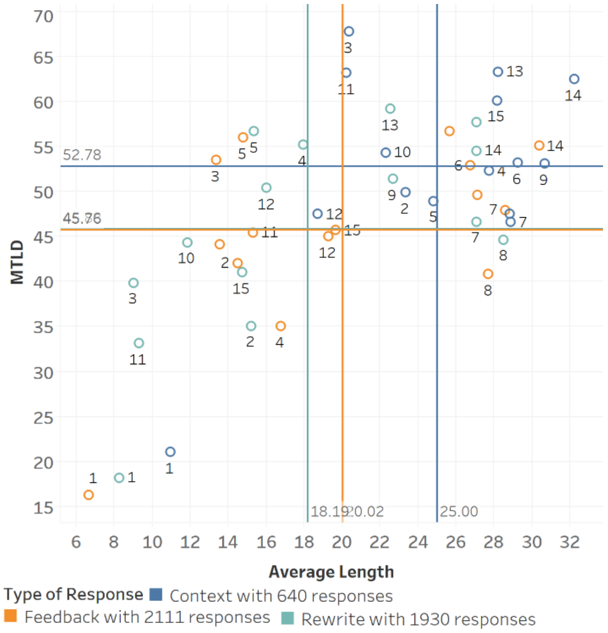

FIGURE III

Plot of MTLD and average response length over all questions by data type (bars give the mean value for each data type, numbers indicate the corresponding question number) N.B the mean MLTD value for rewrite prompts is partially obscured by the value for feedback prompts

Figure IV shows a T-SNE plot of Glove word vector averages for the question shown in Table I. There are clear clusters for the data with labels 001, 010, and 110. There is noticeable separation between labels 101 and 010 and 101. The data does not present a linear structure, so non-linear models will be investigated during the future work on modeling.

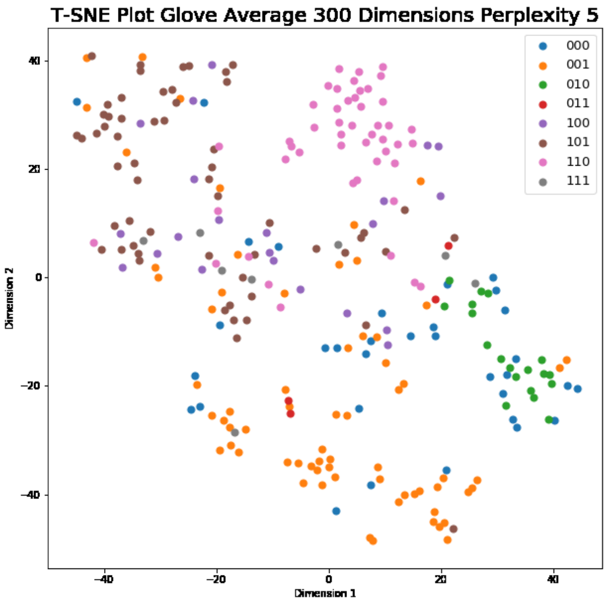

Figure IV

T-SNE plot for the question from Table I (legend uses labels presented in Table II)

II. Algorithm Output

When using the data generation algorithm on sets def(0) and def(1), if there are f_j text snippets in category j for set def(0) and $r_j$in category j for set def(1) there will be $\prod_{j=1}^{n} f_j + r_j$ pieces of data generated from $\sum_{j=1}^{n} f_j + r_j$ pieces of input text. Figure V shows how the output from the data generation algorithm grows as the number of text snippets in the sets increases linearly. The output of the algorithm is $O((\sum_{j=1}^{n} f_j + r_j))^2$for three binary categories and $O((\sum_{j=1}^{n} f_j + r_j))^3$ for five binary categories.

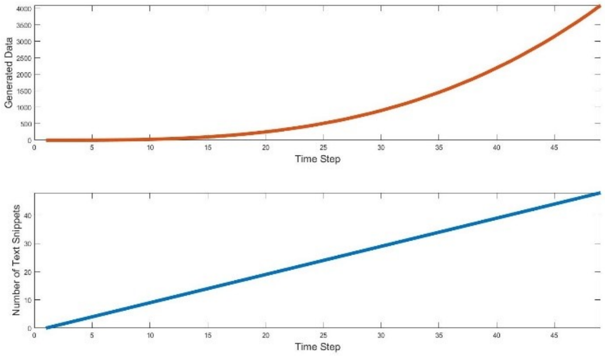

Figure V

PLOT OF COMPARISON OF GROWTH IN TEXT INPUT TO DATA OUTPUT USING THE DATA GENERATION ALGORITHM WITH THREE BINARY CATEGORIES

This shows the growth potential for training data set size when using a data generation scheme based on a binary labeling system.

Conclusion

The three key approaches presented in this paper were the data crowdsourcing method, the binary labeling system, and the data generation system. It was found that context prompts yielded the highest MTLD value and average response length. The binary labeling system offers an approach to extracting annotated features that express the cultural appropriateness of text. This type of annotation approach can be generalized and applied to labeling abstract features like cultural appropriateness or politeness in general human interactions. The binary labeling system makes data generation possible. The presented data generation approach allows for large scale data set creation using a small input data set. The methodology for data collection discussed in this paper can be implemented for replacing a multiple-choice dialogue system with a text classification based free-input dialogue system.

To evaluate the effectiveness of using the synergy of the data crowdsourcing and data generation, the data collection methods presented in this paper will be used to implement a free-input dialogue system in the simulation built by Sheridan et al. [17]. The model will be trained using the training data generated by the presented algorithm. The model will be evaluated on the crowdsourced test data set. The outcome of testing on users of the system will help further evaluate the general accuracy of the classification models used.

A great deal of effort is necessary to collect and label a dataset for building this type free input dialogue system. The data set collected thus far comprises 4,681 labeled pieces of test data. A larger data set may be the key to achieving even better results. Therefore, creating a pipeline for updating the model by adding new data would be an interesting area of further research.

Acknowledgment

We would like to thank the Army Research Lab for supporting this project.

References

[1] Brown, D., Moenning, A., and Guerlain, S., et. al. (2018) "Design and evaluation of an avatar-based cultural training system." *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology.*

[2] Lewis Johnson, W., Vilhjalmsson, H., Marsella, S., et al. (2005). "Serious Games for Language Learning: How Much Game, How Much AI?." *Proceedings of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology* pp. 306-313.

[3] Zielkea M. A., Evans, M. J., and Dufour, F. et. al. (Mar./Apr. 2009) "Serious Games for Immersive Cultural Training: Creating a Living World." *IEEE Computer Graphics and Applications*, vol. 29, no. 2, pp. 49-60

[4] Minkov, M. and Hofstede, G. (2011) "The Evolution of Hofstede's Doctrine." *Cross Cultural Management: An International Journal,* 18(1), pp. 10-20

[5] Hofstede, G. (2011). "Dimensionalizing cultures: The Hofstede model in context." Online Readings in Psychology and Culture, 2(1). Retrieved from doi.org/10.9707/23070919.1014.

[6] Hofstede, G. H., Hofstede, G. J., and Minkov, M. (2010) "Cultures and Organizations." 3rd ed. Maidenhead: McGraw-Hill

[7] Zong, J. and Batalova, J. (2017, September, 29). Chinese Immigrants in the United States. Retrieved from migrationpolicy.org/article/chinese-immigrants-united-states

[8] Behlin, M. (2016, November, 22). "U.S., China participate in Disaster Management" Exchange. Retrieved from army.mil/article/178714/us_china_participate_in_disaster_management_exchange

[9] Wilson , S. R., Kim, M., and Meischke, H. (1991) "Evaluating Brown and Levinson's politeness theory: A revised analysis of directives and face." *Research on Language and Social Interaction,* 25:1-4, pp. 215-252

[10] Huang, Y. (2009). "Politeness Principle in Cross-Culture Communication.", *English Language Teaching. 1.*

[11] Senowarsito, S. "POLITENESS STRATEGIES IN TEACHER-STUDENT INTERACTION IN AN EFL CLASSROOM CONTEXT." *Jalan Sidodadi Timur No. 24, Semarang, Indonesia* pp. 84-96

[12] Bickmore, T., and Cassell, J., (2005) "Social dialog with embodied conversational agents." *In J.C.J. Kuppevelt, L. Dybkjaer, & N.O. Bernson (Eds.), Advances in natural multimodal dialog systems New York: Springer* p 1–32.

[13] Selmeski, Brian R., 16 May 2007. "Military Cross-Cultural Competence: core concepts and individual development." Royal Military College of Canada Centre for Security, Armed Forces & Society Occasional Paper Series—Number 1

[14] Lane, H. C., 2007, "Metacognition and the Development of Intercultural Competence", In Proceedings of the Workshop on Metacognition and Self-Regulated Learning in Intelligent Tutoring Systems at the 13th International Conference on Artificial Intelligence in Education (AIED), pp. 22-33

[15] Roberts, S., Warda, M., Garbutt, S., et. al, 2014, "The Use of High-Fidelity Simulation to Teach Cultural Competence in the Nursing Curriculum", *Journal of professional nursing: official journal of the American Association of Colleges of Nursing.* 30. pp. 259-65.

[16] Taşdemir, S. A., Pasolova-Førland, E., 2014, "Visualizing Afghan Culture in a Virtual Village for Training Cultural Awareness in Military Settings.", *18th International Conference on Information Visualisation*, pp. 256-261

[17] Sheridan, M., An, B., and Brown, D., 27 April 2018, "Investigating the Effectiveness of Virtual Reality for Cross-Cultural Competency Training", *2018 Systems and Information Engineering Design Symposium (SIEDS),* pp. 53-57

[18] Ward, N. G. 2019 "Planning for a Corpus ofContinuous Ratings of Spoken Dialog Quality.", *UTEP Computer Science Technical Report, UTEP-CS* pp. 19-28

[19] Kang, Y., Zhang, Y., Kummerfeld, J. et al. "Data Collection for Dialogue System: A Startup Perspective." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)* pp. 33-40

[20] Serban, I. V., Sankar, C., Germain, M. et al. (2017) "A Deep Reinforcement Learning Chatbot." *ArXiv e-prints*

[21] Viera, A. J., Garrett, J. M., (2005) "Understanding Interobserver Agreement: The Kappa Statistic." *Family Medicine Vol. 37, No. 5* pp. 360-363

[22] Koizumi, R. August 2012 "Relationships Between Text Length and Lexical Diversity Measures: Can We Use Short Texts of Less than 100 Tokens?" *Vocabulary Learning and Instruction Volume 1, Issue 1* pp.60-69

Author Information

Vaibhav Sharma, Graduate Student, Data Science Institute, University of Virginia.

Beni Shpringer, Graduate Student, Data Science Institute, University of Virginia.

Sung Min Yang, Graduate Student, Data Science Institute, University of Virginia.

Martin Bolger, MS Student, Department of Engineering Systems and Environment, University of Virginia.

Sodiq Adewole, Ph. D Student, Department of Engineering Systems and Environment, University of Virginia.

Dr. Don Brown, Professor, Department of Systems and Information Engineering, University of Virginia.

Erfaneh Gharavi, Research Assistant, Department of Engineering Systems and Environment, University of Virginia.