Building and Testing a Chatbot VR Training Simulation

Vaibhav Sharma, Beni Shpringer, Sung Min Yang

Martin Bolger, Sodiq Adewole, Erfaneh Gharavi

Advisor: Donald E. Brown

ARL

University of Virginia

University of Virginia
Data Science Institute

Overview

- Introduction of Problem
- Relevant Work
- Chatbot System Blueprint
- Data Collection and Generation
- Text Classification Overview
- Modeling
- Experiment and Conclusions

Introduction of Problem

Many people in many industries go through behavioral training

Motivating Question:

Can we create a more effective method of behavioral training, using voice-input simulations?

Goal:

Create, and test the effectiveness of, a voice-input simulation, in the context of training US military officers to interact with their Chinese counterparts.

This type of framework could be used for developing a variety of educational simulations.

Relevant Work

Past Work: Simulation-based Training

- Past work at UVa has established the effectiveness of simulation-based cross-cultural competence training, as compared to non-engaging training: pamphlets, powerpoints, etc. [1]
- Previous simulations used multiple-choice input.[2]
- This makes a study of multiple-choice vs. free-response simulation an interesting next step.

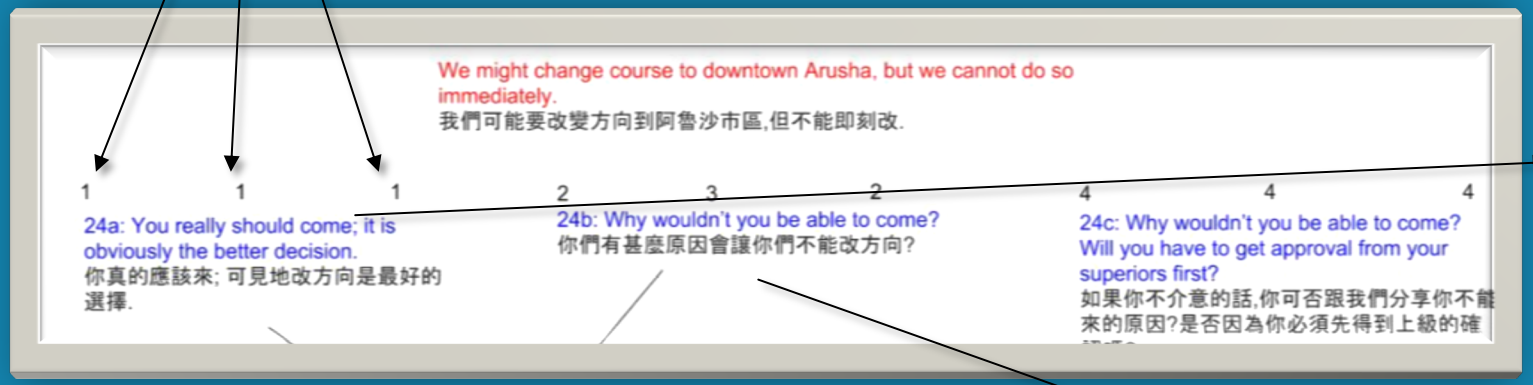

Good morning, Captain Wang. I am honored to have you join us.
Good morning, Captain Wang.
Good morning, Captain Wang. How are you doing today?

Past Work: Existing Simulation

- Context: US Lieutenant (user) coordinating with Chinese Captain (avatar) to assist refugees after a natural disaster.
  - Scenarios were based on real training exercise
- Method: Point-and-Click/Multiple-Choice
- Training Mechanism: Feedback based on dialogue choices

Example Dialogue

User's dialogue options are ranked by three raters

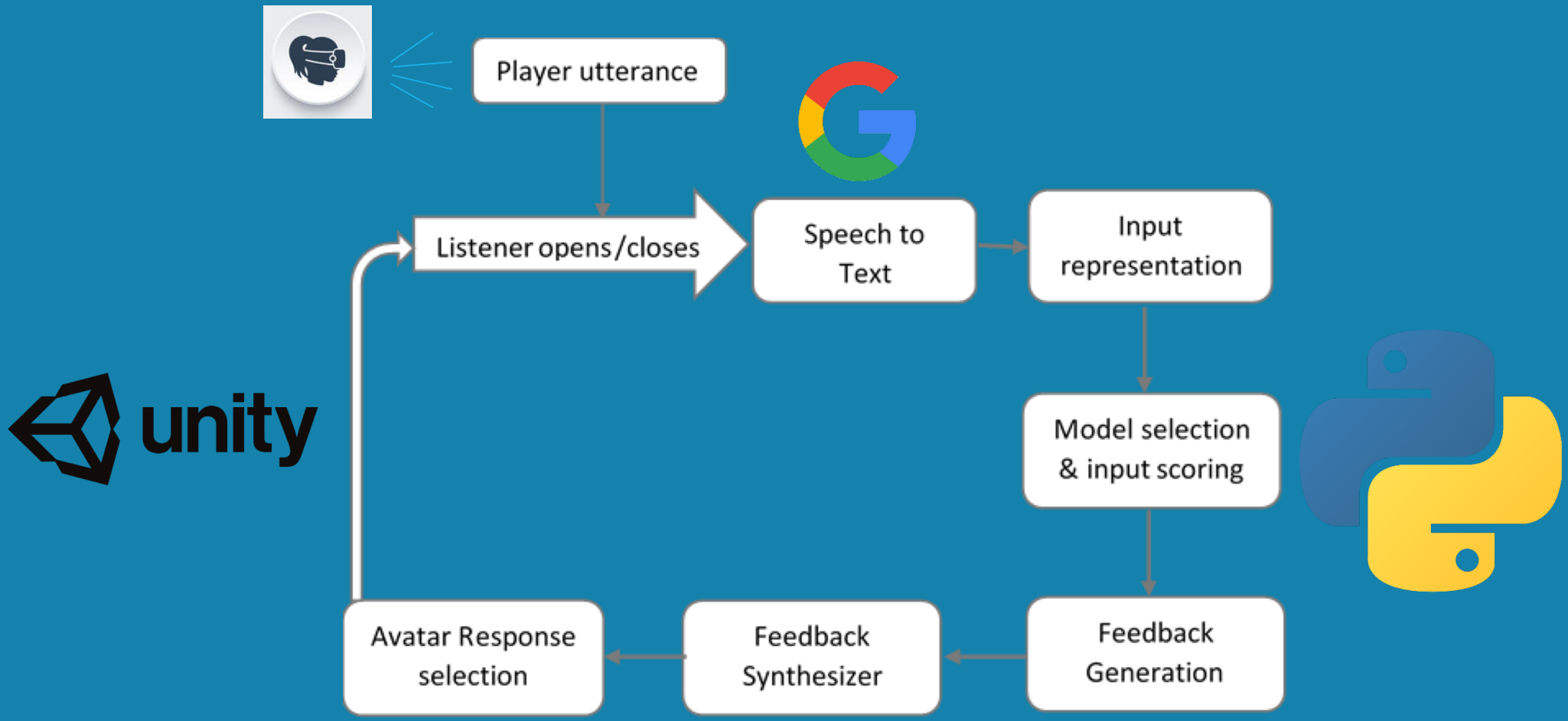


Dialogue	Feedback
You really should come; it is obviously the better decision.	This selection is poor because it comes across as insulting to the intelligence of your interlocutor.
Why wouldn't you be able to come?	This selection is acceptable because it is sensitive to the reality that there are legitimate reasons that may keep the Chinese from changing course. It could be improved by adding a line anticipating that the Chinese may have to first obtain approval for the change from their higher headquarters.
Why wouldn't you be able to come? Will you have to get approval from your superiors first?	This is the best selection because it is simultaneously seeking to understand the Chinese interests while also demonstrating awareness of how Chinese hierarchy works.

- Raters agree on the feedback which explains the reason the dialogue option is correct or incorrect.
- Feedback is displayed in the game after each dialogue selection.

Chatbot System Blueprint

Chatbot System Blueprint



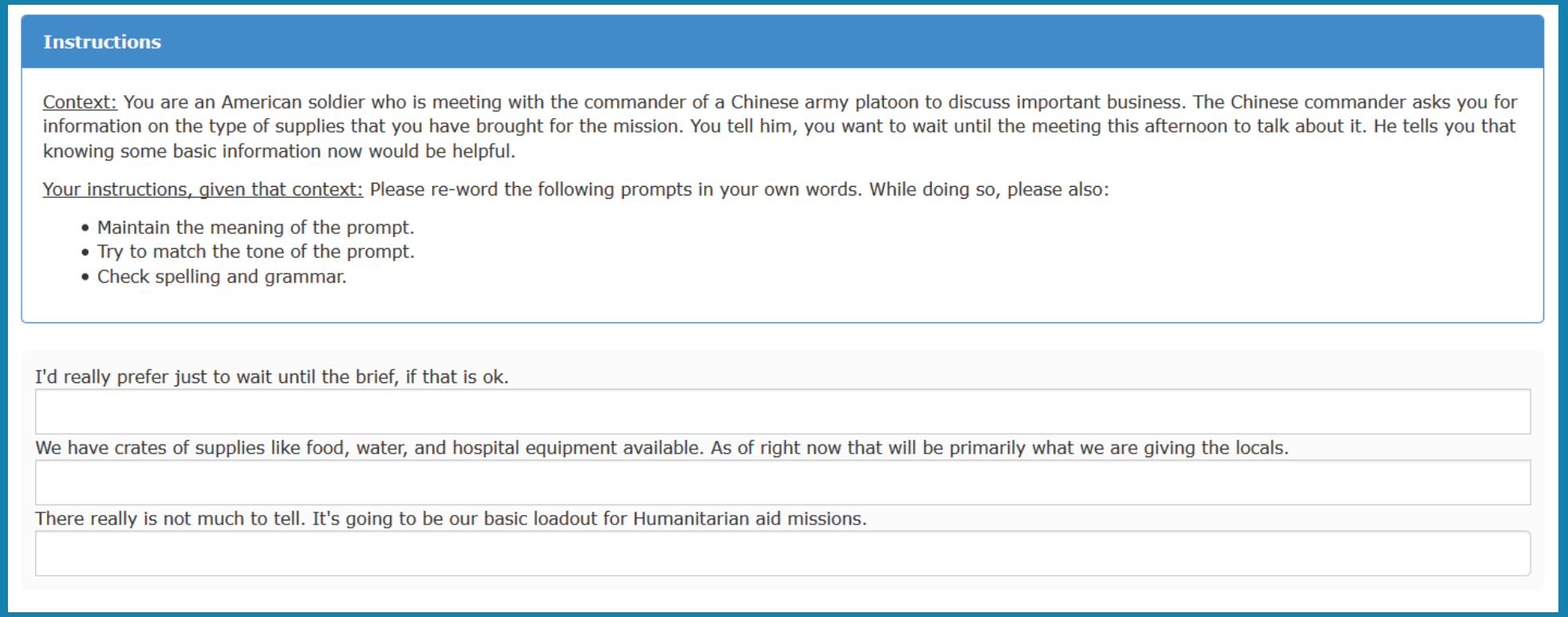
Data Collection and Generation

Data Set Collection

- Original Data Set:

Data Set Collection

- To train a text classifier, we needed a lot of data.
- Data for building the model is collected on Amazon Mechanical Turk
  - Amazon Mechanical Turk is a platform for collecting survey data or labels



Data Set Collection

Amazon Mechanical Turk Responses

Rephrasing

- Data is collected by asking users to rephrase the original multiple-choice options.
- Pre-associated with a label

Feedback

- Data is collected by users creating responses that match the feedback for original the multiple-choice options
- Pre-associated with a label

Context

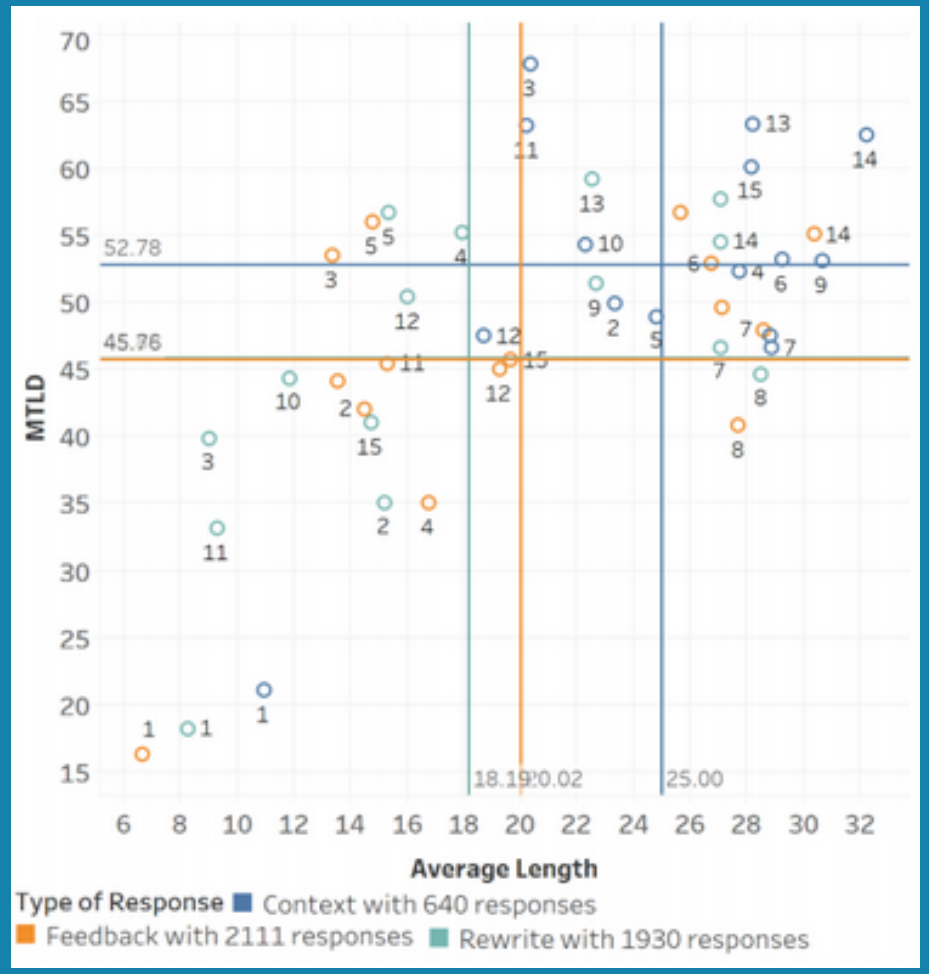
- Data is collected by giving users the context for an in-game question and asking them to create an answer
- Requires labeling/rubric

Data Labeling

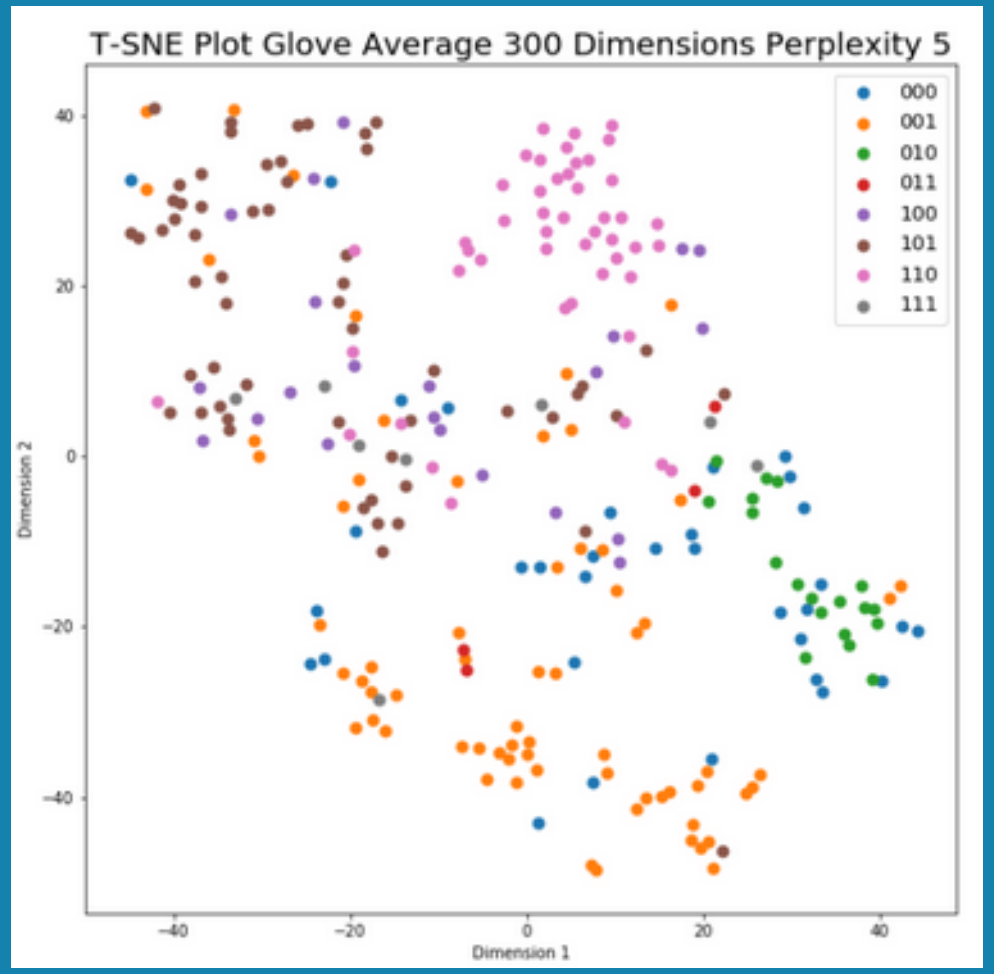
	Gives time information will be given	Displays understanding and empathy	Explains that you are not ready
Is it alright if this waits until the brief? I am currently busy preparing for it.	1	0	1
I appreciate you looking out for your team, but you will get all of it soon when we have the brief this afternoon.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

- The Cohen's Kappa score for the labels was above 0.61 for each question
- This indicates substantial inter-rater agreement

Data Collection Results



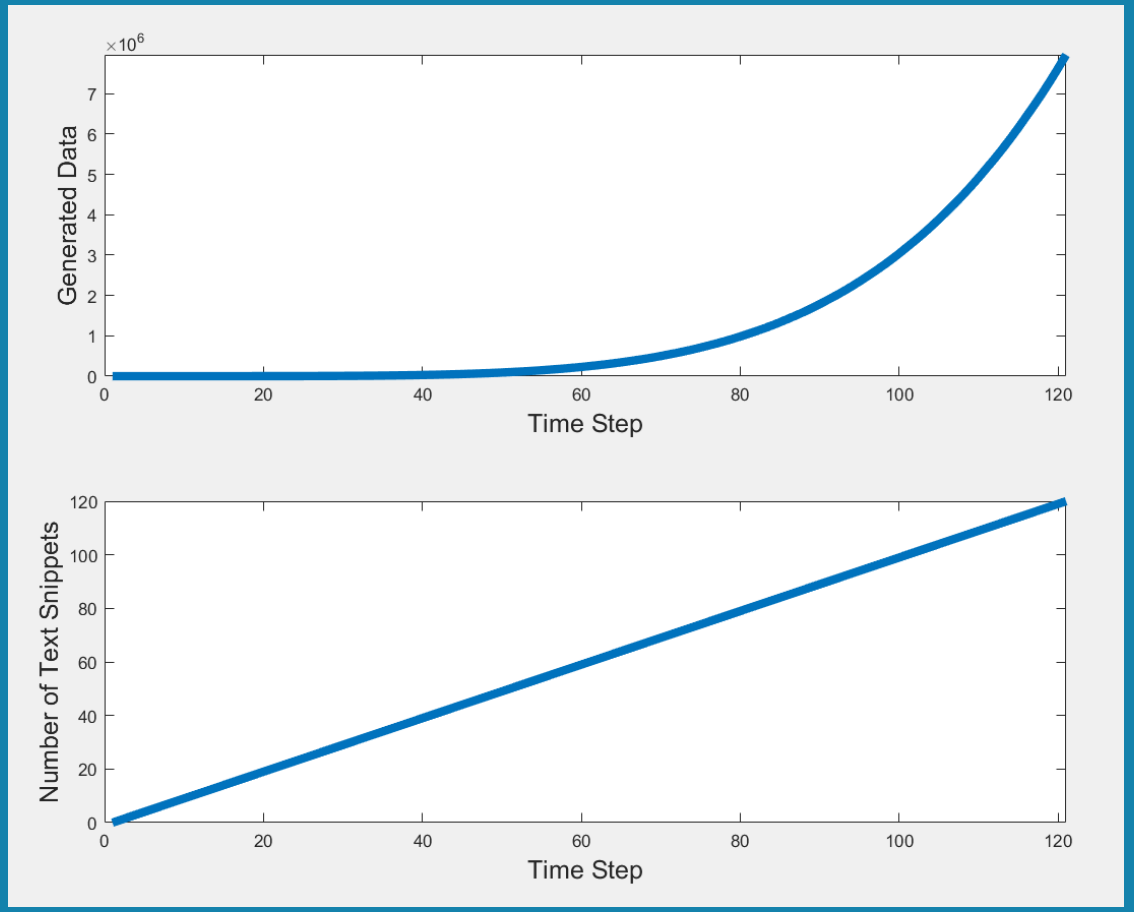
MTLD = Measure of Textual Lexical Diversity, a measure of lexical diversity that is not as sensitive to text length[7]



TSNE plot shows the differences in characteristics for responses with different labels

Data Set Generation

- A data generation algorithm was used to create a larger training set
- Data is generated by creating combinations of example text that matches each binary label
- This creates a large dataset from a small amount of input data
- The patterns in the generated data can be used to help the classifier learn the patterns in input data



Text Classification

Text Feature Representation: Bag-of-Words

Bag-of-Words raw document frequency count Example:

Document 1: “John wants to eat watermelon. Mary wants to eat eggs”

Document 2: “Mary and I like to eat the same things”

Document 3: “John and I are going to eat later”

BoW representation:
D1 = {(“John”, 0) = 1, (“wants”,1) = 2, (‘to’,2) = 2 , (‘eat’, 3)=2, (‘watermelon’,4) =1, (‘Mary’,5) =1, (‘eggs’,6) = 1}
D2 = {(‘Mary,5) = 1 (‘and’, 6) = 1, (‘I’, 7)= 1, (‘like’, 8) = 1, (‘to’, 2) = 1, (‘eat’, 3) = 1, (‘the’, 9)=1, (‘same’, 10) = 1, (‘things’, 11) = 1}
D3 = {(‘John’, 0)= 1, (‘and’, 6)=1, (‘I’, 7) = 1, (‘are’ = 12) = 1, (‘going’, 13)=1, (‘to’, 2) = 1, (‘eat’, 3) = 1, (‘later’, 14)=1}

Term Frequency

	Document 1	Document 2	Document 3
(John, 0)	1	0	1
(Likes, 1)	2	0	0
(to, 2)	2	1	1
3	2	1	1
4	1	0	0
5	1	1	0
6	0	1	1
7	0	1	1
8	0	1	0
9	0	1	0
10	0	1	0
11	0	1	0
12	0	0	1
13	0	0	1
14	0	0	1

TF-IDF

	Document 1	Document 2	Document 3
(John, 0)	Log(3/2) ≈0.176	0	Log(3/2) = 0.176
(Likes, 1)	Log(3/2)*2 ≈ 0.352	0	0
(to, 2)	Log(1)*2 = 0	Log(1) = 0	Log(1) = 0
3	Log(1)*2 = 0	Log(1) = 0	Log(1) = 0
4	Log(3) ≈ 0.477	0	0
5	Log(3/2) ≈ 0.176	Log(3/2) ≈ 0.176	0
6	0	Log(3/2) ≈ 0.176	Log(3/2) ≈ 0.176
7	0	Log(3/2) ≈ 0.176	Log(3/2) ≈ 0.176
8	0	Log(3) ≈ 0.477	0
9	0	Log(3) ≈ 0.477	0
10	0	Log(3) ≈ 0.477	0
11	0	Log(3) ≈ 0.477	0
12	0	0	Log(3) ≈ 0.477
13	0	0	Log(3) ≈ 0.477
14	0	0	Log(3) ≈ 0.477

- For the i, j^{th} term, we get $weight_i = \log(\frac{N}{\sum_{j=1}^{N} u(t_i \in d_j))})$
- where N is the total number of documents and $\mathrm{u(x, y)} = \begin{cases} 1 & x \in y \\ 0 & otherwise \end{cases}$
- Now $x_{i,j} = n(t_i \in d_j) weight_i$

Word Vectors

- A word vector is a row of real valued numbers (as opposed to dummy numbers) where each point captures a dimension of the word's meaning and where semantically similar words have similar vectors
- So words like king and queen will have similar dimensions in the vector space
- Used approaches like CBOW and skip gram to make them and used a soft-max function to generate probability distribution over the output classes

$$w_1 \quad w_2 \quad w_3 \quad \dots \quad w_n$$

$$\begin{bmatrix}1\\0\\0\\\vdots\\0\\0\\0\end{bmatrix} \quad \begin{bmatrix}0\\1\\0\\\vdots\\0\\0\\0\end{bmatrix} \quad \begin{bmatrix}0\\0\\1\\\vdots\\0\\0\\0\end{bmatrix} \quad \dots \quad \begin{bmatrix}0\\0\\0\\\vdots\\0\\0\\1\end{bmatrix} \in \mathbb{R}_{nx1}$$

Modeling Results

Best Modeling Results

Modeling Results on 200 Dim TFIDF

Question	Binary Classes	Total Classes	RF	KNN	MLP	CNN	LSTM	GRU
1	3	8	0.73	0.78	**0.86**	0.81	0.75	0.71
2	3	8	0.31	0.45	0.56	0.55	0.37	0.42
3	2	4	0.56	0.55	**0.73**	0.72	0.52	0.24
4	3	8	0.32	0.54	0.59	**0.67**	0.35	0.46
5	2	4	**0.88**	0.78	**0.90**	0.89	0.50	0.23
6	3	8	0.76	0.71	0.76	**0.79**	0.61	0.68
7	3	8	0.70	**0.75**	0.69	0.71	0.46	0.46
8	4	16	0.48	0.59	0.58	0.44	0.40	0.03
9	3	8	0.36	0.49	0.56	**0.57**	0.12	0.08
10	2	4	0.83	0.80	**0.86**	0.57	0.00	0.00
11	1	2	**0.85**	0.80	**0.85**	N/A	N/A	N/A
12	2	4	0.79	0.73	**0.80**	0.54	0.11	0.17
13	3	8	0.41	0.48	**0.52**	0.50	0.20	0.18
14	2	4	**0.65**	0.60	0.64	0.46	0.10	0.10

Finished Product


Good morning, Captain Wang. I am honored to have you join us.
Good morning, Captain Wang.
Good morning, Captain Wang. How are you doing today?


Guide |It's morning, greet Captain Wang

Design of Experiment

Design of Experiment Overview

Questionnaire:
Statistics Collection + Innate Performance Predictor E-CQS

Control:
Old Simulation (Multiple Choice)

Target:
New Simulation (Free-Input)

Earthquake Scenario Based Post Test

Hypothesis and Main Metric

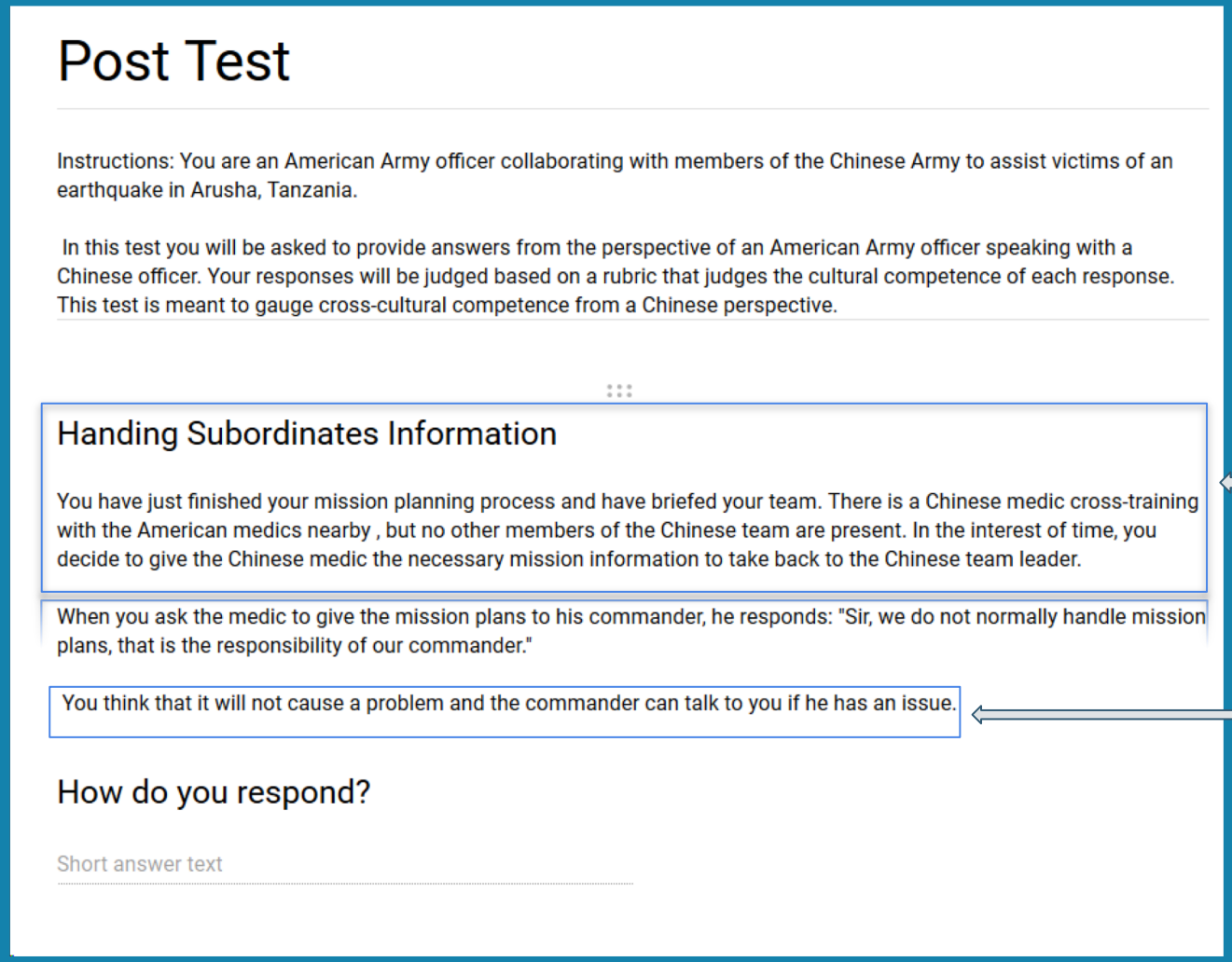
The Hypothesis:

Subjects trained using the new free-input simulation score higher on our test of cross-cultural competency than subjects trained using the old multiple choice simulation. Accept hypothesis if there is statistically significant improvement in the average of taker's post-test paper test scores (0.05 significance level)

The Main Metric:

The subject's score on the post test is the main metric for evaluating the hypothesis. The data collected in the questionnaire and E-CQS will be used to control for each subject's possible inherent biases.

Post Test Example



Post Test

Instructions: You are an American Army officer collaborating with members of the Chinese Army to assist victims of an earthquake in Arusha, Tanzania.

In this test you will be asked to provide answers from the perspective of an American Army officer speaking with a Chinese officer. Your responses will be judged based on a rubric that judges the cultural competence of each response. This test is meant to gauge cross-cultural competence from a Chinese perspective.

Handing Subordinates Information

You have just finished your mission planning process and have briefed your team. There is a Chinese medic cross-training with the American medics nearby , but no other members of the Chinese team are present. In the interest of time, you decide to give the Chinese medic the necessary mission information to take back to the Chinese team leader.

When you ask the medic to give the mission plans to his commander, he responds: "Sir, we do not normally handle mission plans, that is the responsibility of our commander."

You think that it will not cause a problem and the commander can talk to you if he has an issue.

How do you respond?

Short answer text

Question: Explains the situation. The header indicated the skill the question is testing

Motivation: Explains the character's motivation in this situation. This will guide the test taker to the right type of answer.

Post Test Example

Post Test

Instructions: You are an American Army officer collaborating with members of the Chinese Army to assist victims of an earthquake in Arusha, Tanzania.

In this test you will be asked to provide answers from the perspective of an American Army officer speaking with a Chinese officer. Your responses will be judged based on a rubric that judges the cultural competence of each response. This test is meant to gauge cross-cultural competence from a Chinese perspective.

Handing Subordinates Information

You have just finished your mission planning process and have briefed your team. There is a Chinese medic cross-training with the American medics nearby , but no other members of the Chinese team are present. In the interest of time, you decide to give the Chinese medic the necessary mission information to take back to the Chinese team leader.

When you ask the medic to give the mission plans to his commander, he responds: "Sir, we do not normally handle mission plans, that is the responsibility of our commander."

You think that it will not cause a problem and the commander can talk to you if he has an issue.

How do you respond?

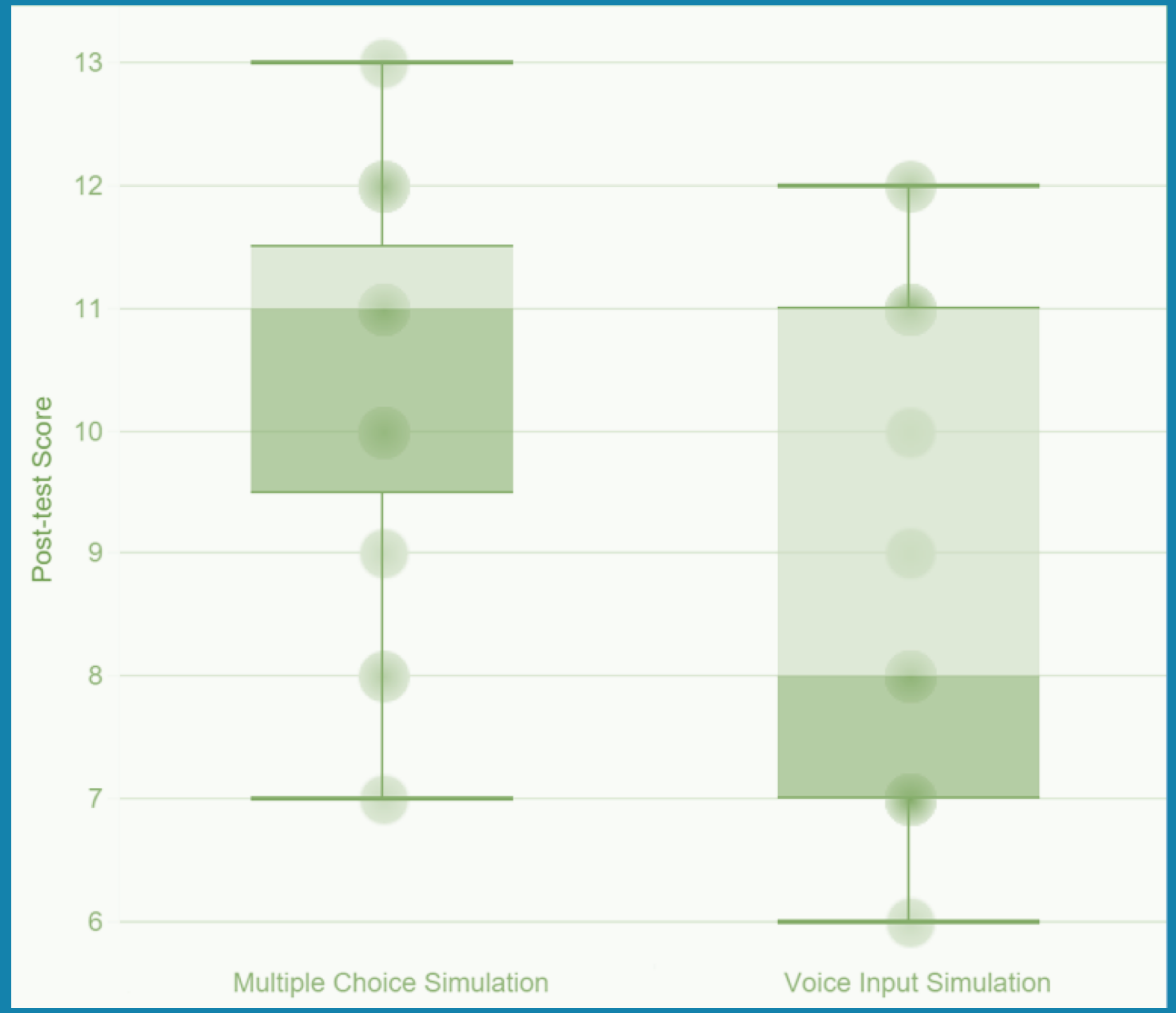
Short answer text

Rubric for question:

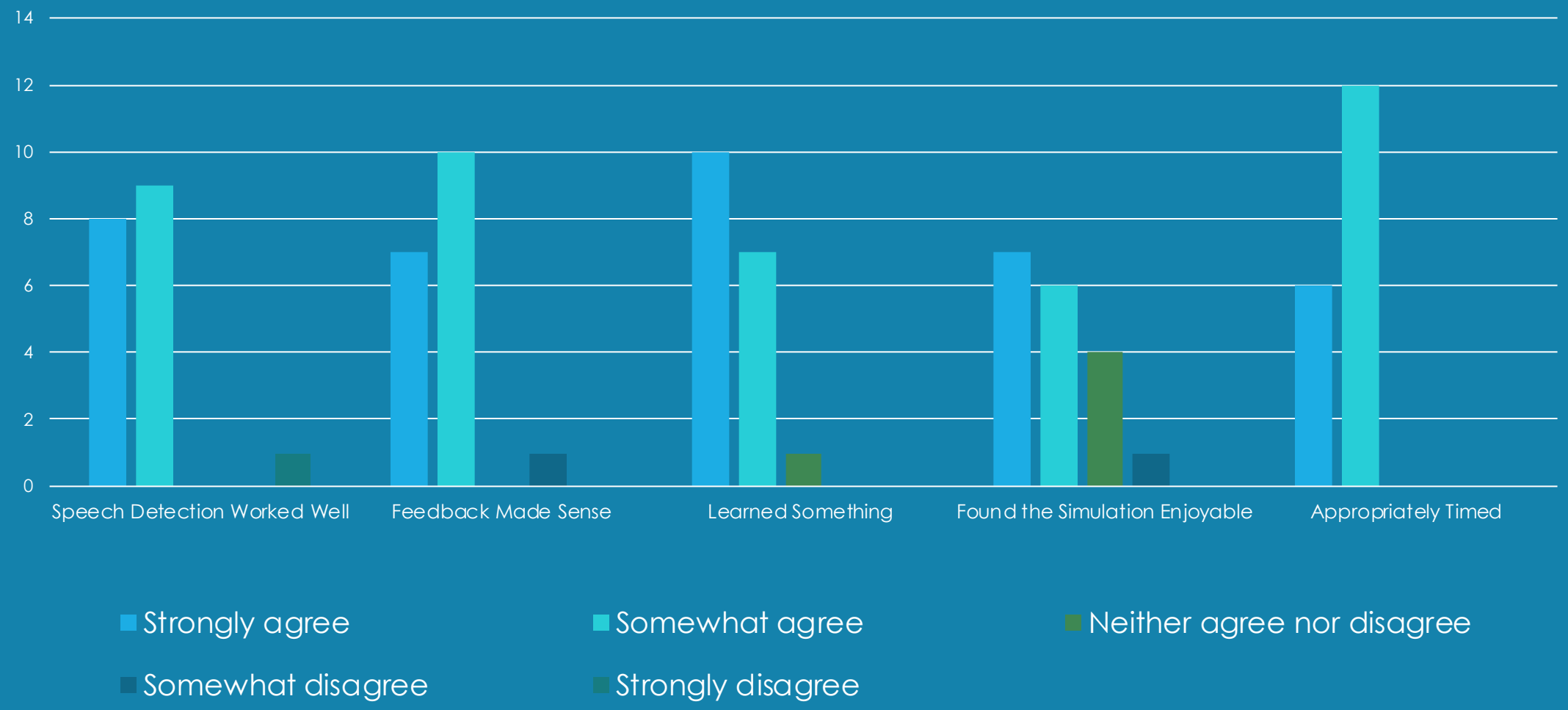
- Good response: Frames request as beneficial to the medic's superior's interests.
- Average response: Dismissive of the Chinese medic's concerns.
- Poor response: Defers responsibility to the Chinese team and asks commander to do something.

Experiment Results and Conclusions

- Given the small sample size (N=35), the experimental hypothesis currently cannot be rejected
- A larger scale test should be considered to definitively come to a conclusion about the effectiveness of these training methods.
- We'd like to further test future-recall of lessons learned from the training process, as this is more relevant to understanding the efficacy of this kind of training. That is, how well remembered are the lessons weeks or months after the training.



User Experience Survey Results



Impact

Impact

- Developed a method create a data set for a specific scenario/simulation, which can be used for text-classification.
- We found methods/algorithms that are capable of accurately classifying speech, in order to provide feedback.
- We showed that people can enjoy this kind of training through our testing.
- We have set the ground work for more testing regarding the impact of this sort of training method on learning.

QUESTIONS?

THANK YOU!

References

1.) Sheridan, Martha et al. "Investigating the effectiveness of virtual reality for cross-cultural competency training SIEDS 2018." 2018 Systems and Information Engineering Design Symposium (SIEDS) (2018): 53-57.

2) 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

3.) Johnson, W & Högni Vilhjálmsson, Hannes & Marsella, Stacy. "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 (2005). 306-313.

4.) Johnson, W.L., Marsella, S., Mote, N., Viljhalmsson, H, Narayanan, S., Choi, S. "Tactical Language Training System: Supporting the rapid acquisition of foreign language and cultural skills." InSTIL/ICALL Symposium, Venice, Italy. (2004).

5.) Zielke M. A., Evans M. J., Dufour F., Christopher T. V., Donahue J. K., Johnson P., Jennings E. B., Friedman B. S., Ounekeo P. L., Flores R., "Serious Games for Immersive Cultural Training: Creating a Living World." IEEE Computer Graphics and Applications, vol. 29, no. 2, (2009) pp. 49-60, Mar./Apr.

6.) Kron FW, Fetters MD, Scerbo MW, White CB, Lypson ML, Padilla MA, et al. "Using a computer simulation for teaching communication skills: A blinded multisite mixed methods randomized controlled trial." Patient Educ Couns (2016)

7.) I. NEWBLE, D et. al. "A comparison of multiple choice and free response tests in examinations of clinical competence. Medical Education." (1979). 13. 263 - 268

8.) 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

9.) D'Souza, Jocelyn (2018, April 4), *An Introduction to Bag-of-Words in NLP* retrieved from https://medium.com/greyatom/an-introduction-to-bag-of-words-in-nlp-ac967d43b428

10.) Chaubard, Francois et al. (2016) *CS224D: Deep Learning for NLP Lecture Notes: Part I* retrieved from https://cs224d.stanford.edu/lecture_notes/notes1.pdf