Examining Linguistic Behavior of a Virtual Museum Guide

Undergraduate Research Thesis

Presented in partial fulfillment of the requirements for graduation *with honors research distinction* in Linguistics in the undergraduate colleges of The Ohio State University

by

Madeleine Alette Bloomquist

The Ohio State University

May 2023

Project Advisor: Professor William Schuler, Department of Linguistics

Introduction

As artificial intelligence that uses natural language processing (NLP) becomes more prevalent, analyzing such software so that it maximally understands people holds all the more significance. While recurrent neural networks can predict the next word in a sentence with high accuracy, dialogue systems do not engage with humans as humans do with each other. Despite the robust development of NLP technologies, their capabilities are not meaningful if those who interact with them do not find them to be practical. A human speaker can produce an unlimited variety of stimuli, and variability between speakers is immense. The response that such stimuli yields from dialogue systems reveals how NLP software handles variability. Evaluating its errors and searching for patterns among them provides insight as to where software can be improved.

At the Center of Science and Industry (COSI), a science museum in Columbus, Ohio, researchers from The Ohio State University have developed an interactive avatar that uses NLP. In this context, an avatar can be defined as a human-like bot created to interact with users. Visitors can ask the avatar questions related to linguistics, computer science, and exhibits at the museum. The avatar consists of both an animated visual component and its artificial intelligence software, which processes speech as input and produces a response accordingly. In this case, the artificial intelligence used to process language is a self-attentive recurrent neural network, trained on a corpus of English text pertaining to computer science, linguistics, and COSI exhibits. This research focuses on the effectiveness of the avatar's responses to human user input. In calculating accuracy, a group of unanswerable user queries was identified; these were given special consideration in later reanalysis.

Background

Virtual museum guides have been implemented as a strategy to increase public engagement in various large science museums throughout the world, including the Museum of Science, Boston (MoS) and the Heinz Nixdorf Museums Forum (HNF) in Paderborn, Germany (Swartout et al. 2010), as well as the Center of Science and Industry (COSI) in Columbus, Ohio (Maicher et al. 2019). Lifelike visuals combined with communication in natural language make these virtual humans as engaging to visitors as actual human guides. Unlike the virtual humans in use at, for example, the HNF, the avatar at COSI is designed to interact directly with museum guests in a spoken modality. Rather than asking questions to a handler or typing input via keyboard, visitors at COSI speak directly to the avatar (Swartout et al. 2010). The avatar responds verbally to these queries. Despite the dialectal variation among guests, the background noise inherent to a crowded public space, and the numerous ways that the same question can be worded differently, the avatar is still able to process this input thanks to the sophistication of natural language technologies.

The system that is currently in use at COSI was originally based on a previously developed Virtual Standardized Patient (VSP), a simulated character that acts as a sick patient in order to give medical students practice taking medical histories (Maicher et al. 2019). The VSP uses the a combination of rule-based and data-driven methods to take human speech, transform it into a medium that it can process, and generate a response that fits in the context of the conversation (Stiff et al. 2022). The avatar at COSI functions similarly, using automated speech recognition (ASR) to encode user input, identifying the question with artificial intelligence (AI), and responding with automated speech generation (Maicher et al. 2022). The self-attentive recurrent neural network in the AI step uses word embeddings to make predictions about

language based on a corpus of English text in the training process (Sunder & Fossler-Lussier, 2021). The recurrent neural network is trained on text, in this case that is specific to the museum, linguistics, and computer science. Pattern rules are created from the predictions it makes in the training process. The meaning of each word in a given sentence is determined by the context of the surrounding words. Neural networks can be conceptualized in terms of observation vectors and context vectors. The context vectors are connected to the vector for the current word, and in models that use attention, an attention matrix is formed, assigning weights of importance to input words. Attention models generate output based on the weighted importance of all given input. Once word representations are understood in context, attention is used to classify user input. As more training data is used, the model gets better at understanding natural language. Whether or not it has been exposed to a particular sentence before, it can glean a meaning given the words that it already understands.

In addition to this natural language processing technology, the Unity game engine is used to create the visual component of the avatar: a three-dimensional, virtual human woman in a lab coat named Dr. Lehiste. Dr. Lehiste appears on the screen of an iPad in the Language Pod at COSI, a designated area to conduct linguistic research with those who wish to participate. Over 150 museum visitors have participated in conversations with the avatar since its creation.

The avatar matches the guest's utterance with the closest entry in its bank of questions to which it has answers, then responds with said answer. Included in these answers are also a response asking to rephrase the question, as well as topic suggestions for low-confidence questions. The weighted attention mechanism introduced above is used to classify questions.This research began as a project to determine the avatar's accuracy of responses to users' questions, and later additionally focused on how the avatar categorized users' questions.

Method

The avatar usage data that was received from COSI came in the form of comma-separated values (CSV) files, containing participant demographics, transcriptions of participant utterances, the canonical questions that the system matched to them, and the responses the avatar gave, among other data points. For the first phase of the project, the accuracy of the avatar was examined, which has been defined as whether or not the response it gave answered the participant's question. A spreadsheet was created in which each participant was numbered, as were the number of turns in each conversation. In one column, the accuracy of each turn was assessed, marked with a Y for "yes, accurate" or an N for "no, not accurate." Accuracy was determined by comparing the transcription of the participant's utterance to the question that the avatar classified it as, then making a judgment on whether or not the avatar's classification was correct. For example, if a participant asked, *Where are the bathrooms?* and the avatar matched it with the question in its bank, *Where is the nearest bathroom?* it would be marked as accurate. Conversely, if a participant asked, *How old are you?* and the avatar classified it as, *How are you?* it would be marked as inaccurate. Participant identity was also tracked as it was written in the CSV file, accuracy as a percentage for each participant, and the overall accuracy as a percentage for all of the participants whose data was used.

After this process was completed with 31 participants, the participant utterances to which the avatar responded inaccurately were examined. While the hope was to find a syntactic pattern in the participant utterances with which the avatar struggled, it was found that most of the inaccurate responses were to non-questions, utterances that were not addressing the avatar, or questions that were outside of the scope of the avatar's knowledge. Examples of these types of input from the data include *I got two questions here*, a non-question; *Yeah, ask your question*,

which is not addressing the avatar; and *Where did we come from?* which is out of scope. Changing the approach, the instances that were out of scope or not addressing the avatar were then filtered out. Adjusted accuracy was added for each participant and for all of the participants overall. Any avatar responses asking to rephrase the question were considered accurate.

Once the filtering process was completed for all 31 participants, correct labels were annotated for the inaccurate responses. In a new spreadsheet, the number in the CSV file for each turn, transcriptions of the participant utterances, canonical questions that the avatar matched to them, the better match selected from the canonical questions, and the label number for the latter canonical question were tracked. Finding the participant utterances and canonical questions to which the avatar matched them was a straightforward process. In order to find the canonical questions that fit the utterances better, a master document that contained all of the possible canonical questions for which the avatar had a response was searched. Each had a label number, which was documented in the spreadsheet along with the canonical question itself. Utterances that were not addressing the avatar or asking out-of-scope questions were not given a corrected label.

This process for finding accuracy and making label corrections was continued with an additional 39 participants after the original 31, for a total of 70 participants. Participants were a combination of COSI visitors, as well as research assistants and volunteers involved with the Language Pod and the Virtual Patient project.

A later analysis was conducted, giving special consideration to the avatar's responses to input that was not addressing it. Given that the avatar's design was originally intended for the Virtual Standardized Patient, used in a setting in which only one speaker at a time would interact with the dialogue system, addressee detection was not a factor under consideration during its

creation. Therefore, the best response that the avatar could give to input that was not addressing it was a request to rephrase, similar to the best response it could give to out-of-scope questions. Using this new criteria, input that did not address the avatar was examined separately, considering rephrase requests as the most accurate response.

Results

The data from all 70 participants was combined, and the avatar's overall accuracy was found to be 46.8%. When the irrelevant utterances were filtered out, the recalculated accuracy was 73.4%.

By the end of the study, 411 labels were corrected from the system's original output, of which 200 were true corrections, 201 were *out of scope*, meaning that the question was not on a topic that the avatar was trained to understand, or *crosstalk*, meaning that the input did not address the avatar, and 10 were questions that did not have a response but were on-topic enough that new suggested labels were written for them. These were used by the programmers involved with the COSI avatar for purposes beyond the scope of this study.

Examining the input that was not addressing the avatar, or *crosstalk,* there were 244 total utterances found that were categorized as such. Of these utterances, the avatar responded with rephrase requests 57 times, or 23.4% of the time. Five times, the avatar classified this input as matching the *you already said that* label in its question bank. The avatar classified this input as the question, *What all do you know?* 15 times. The rest of the utterances that were not addressing the avatar were classified as various other questions within the avatar's question bank that, unlike those mentioned, did not yield a clear pattern.

Discussion

A substantial proportion of the avatar's inaccurate responses were to participant utterances that were not addressing it or not asking questions within its scope of knowledge. While asking to rephrase is the best that the avatar can do for out-of-scope questions, the issue of addressee detection arises in this case, as well as in the case of voice assistants and related technologies.

Addressee detection is problematic for voice assistants due to the multiparty nature of their interactions with humans, which is recognized by linguists, computer scientists, and others who work with such technology (Akhtiamov et al. 2017). The typical solution that is chosen is to use a specific "wake-word" to address the voice assistant, but linguists such as Oleg Akhtiamov and Ingo Siegert argue that having to use wake-words makes the interaction unnatural (Siegert 2021). In the interest of making interactions more naturalistic, several strategies have been implemented and tested.

In one study examining video and audio recordings of human-human-computer interactions, as well as transcripts produced by ASR, various classifiers were tested by the researchers, using unweighted average recall as the criterion for evaluation (Akhtiamov et al. 2017). They found that a meta-classifier combining acoustic, lexical, and syntactic analysis outperformed all other models included in the study (Akhtiamov et al. 2017).

In a later study, fully connected neural networks and long short-term memory (LSTM) models were applied to the aforementioned recordings and transcripts (Pugachev et al. 2017). Their deep neural network model outperformed their bidirectional LSTM model when evaluated based on average recall, although they speculated that a bidirectional LSTM model may perform better in specific conversational contexts (Pugachev et al. 2017).

In a later experiment with several of the same researchers, a baseline was established by comparing unweighted average recall of a spoken dialogue system in a conversation with two humans to that of a child in a conversation with two adults (Akhtiamov et al. 2019). Testing a linear support vector machine and two neural networks under a variety of conditions, they found that mixup, a technique for data augmentation that works agnostic of domain, benefited neural networks, which also outperformed the linear model, yet linear classifiers did not benefit from mixup (Akhtiamov et al. 2019).

Although addressee detection is not strongly associated with one specific subdiscipline of linguistics, it is a vital component of natural language understanding. The increasing complexity and ubiquity of voice assistants has given more recent rise to research in optimizing addressee detection. In this study, the interactive avatar responded to input that was not addressing it one-third of the times that it responded inaccurately. Implementation of an addressee detection mechanism could improve the avatar's accuracy by preventing it from responding to input that is not directed at it. One potential strategy for this could be to train a classifier on the participant input that was annotated as *crosstalk* (input not addressing the avatar), giving it the opportunity to learn what type of input does not necessitate a response.

References

Akhtiamov, O., Siegert, I., Karpov, A., and Minker, W. 2019. Cross-Corpus Data Augmentation for Acoustic Addressee Detection. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 274–283, Stockholm, Sweden. Association for Computational Linguistics.

Maicher, K. *et al.* (2019) Using virtual standardized patients to accurately assess information gathering skills in medical students. In: Zimmerman, L., Wilcox, B., Liston, B., Cronau, H., Macerollo, A., Jin, L., Jaffe, E., White, M., Fosler-Lussier, E., Schuler, W., Way, D., Danforth, D. Medical Teacher, 41:9, 1053-1059, DOI: 10.1080/0142159X.2019.1616683

Maicher, K. *et al.* (2022) Artificial intelligence in virtual standardized patients: Combining natural language understanding and rule based dialogue management to improve conversational fidelity. In: Stiff, A., Scholl, M., White, M., Fosler-Lussier, E., Schuler, W., Serai, P., Sunder, V., Forrestal, H., Mendella, L., Adib, M., Bratton, C., Lee, K., & Danforth, D. Medical Teacher, 45:3, 279-285, DOI: 10.1080/0142159X.2022.2130216

Pugachev, A., Akhtiamov, O., Karpov, A., Minker, W. (2018). Deep Learning for Acoustic Addressee Detection in Spoken Dialogue Systems. In: Filchenkov, A., Pivovarova, L., Žižka, J. (eds) Artificial Intelligence and Natural Language. AINL 2017. Communications in Computer and Information Science, vol 789. Springer, Cham. https://doi.org/10.1007/978-3-319-71746-3_4

Stiff, A., White, M., Fosler-Lussier, E., Jin, L., Jaffe, E., & Danforth, D. (2022). A randomized prospective study of a hybrid rule- and data-driven virtual patient. *Natural Language Engineering,* 1-42. doi:10.1017/S1351324922000420

Siegert, I., *et al.* (2021) Admitting the Addressee Detection Faultiness of Voice Assistants to

Improve the Activation Performance Using a Continuous Learning Framework. In: Akhtiamov, O., Kruger, J., Weisskirchen, N., Wendemuth, A. *Cognitive Systems Research*, vol. 70, 2021, pp. 65–79., https://doi.org/10.1016/j.cogsys.2021.07.005.

Sunder, V., & Fosler-Lussier, E. (2021) "Handling Class Imbalance in Low-Resource Dialogue Systems by Combining Few-Shot Classification and Interpolation," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 7633-7637, doi: 10.1109/ICASSP39728.2021.9413405.

Swartout, W. *et al.* (2010). Ada and Grace: Toward Realistic and Engaging Virtual Museum Guides. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds) Intelligent Virtual Agents. IVA 2010. Lecture Notes in Computer Science(), vol 6356. Springer, Berlin, Heidelberg.