Researchers use Referential Communication Task (RCT) to assess the potential of dementia. We used the RCT in which one participant orally instructs the other to identify an abstract image from four in each round.

While previous studies practiced robot-mediated RCT, the robots usually provide naive feedback. Therefore, we developed a novel dialogue robot based on word embedding and other natural language processing (NLP) techniques to generate natural feedback in RCT and achieve a mutual theory of mind (ToM) between humans and the robot.

Why do we use word embedding? Word embedding numerically captures the semantic relations between words. We wish the robot could show her understanding of the human speaker by replying with alternative words to describe the images. We assume the participants would feel the conversation was natural because the feedback implies the robot’s understanding of the human’s mental state.

How do we generate word embedding? Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language representation model. We further pre-trained, fine-tuned, and validated it on 2,384 transcripts from 96 students describing 48 abstract images. Word embeddings of all tokens are extracted from the fine-tuned model by summing up its last four encoder layers' hidden states or only the last layer's hidden states.

How does the system work? First, shape/object keywords in the input sentence are identified. Next, the system finds each keyword's three nearest shape/object words by calculating the similarities between their embeddings. The nearest words are then used to fill in the feedback templates.

Transcripts are divided into two parts: training BERT and simulating new participants’ making. We measure the feedback’s coherence and relevance to the conversation by the proportion of all feedback that contains words used to describe the image by other participants (referred to as the match ratio thereafter).

We evaluate the performance by calculating the match ratio under different settings. Variations are the type of BERT models and the token representations.

Model variation: Fine-tuned (BERT-FIT). Within task pre-trained and fine-tuned (BERT-ITPT-FIT).

Token-representation variation: Sum of the last four layers’ hidden states. Last layer’s hidden state.

Results: Bootstrapped BERT-FIT and BERT-ITPT-FIT are cross-validated on the image classification task. No significant difference (p=0.5815) was found between the classification accuracy of BERT-FIT (m=0.8186, s=0.0318) and BERT-ITPT-FIT (m=0.8149, s=0.0293).

The match ratio (at least 82%) in Figure 2 depicts that all of the models can generate coherent and relevant feedback for the ongoing conversation. Compared to extracting the embedding from the last encoder layer, summing up the last four layers enables the model to generate more relevant feedback. However, pretraining does not significantly improve performance. The limited data size might cause its inefficiency.

References:
Peng, R., Liu, Z., Yuan, F., Zeng, M., Zhao, X., Pasquarelle, R. J. A database of multimodal data to construct a simulated dialogue partner with varying degrees of cognitive health p. 8

This project was sponsored by the National Science Foundation through Research Experience for Undergraduates (REU) award and the National Institute of Computational Sciences at University of Tennessee Knoxville.

Acknowledgements.

The University of Tennessee Knoxville