Generating Natural Language Responses in Robot-mediated Referential Communication Tasks

Students: Yigang Qin, Huiqi Zou

Mentors: Ziming Liu, Dr. Xiaopeng Zhao, Dr. Kwai Wong

Abstract

With advances in neural network-based computation, socially assistive robots (SARs) have been endowed with the ability to provide natural conversation to users. However, the lack of transparency in the computation models results in unexpected robots behaviors and feedback, which may cause users to lose their trust in the robot. Theory of mind (ToM) in cooperative tasks has been considered as a key factor in understanding the relationship between user acceptance and the explainability of robot behaviors. Therefore, we develop a dialog system based on word embedding and other natural language processing (NLP) techniques to simulate natural feedback. The system is designed based on the mechanism of ToM and validated with two validation tests. Based on the result, we believe the designed dialog system bears the feasibility of simulating ToM and can be used as a research tool for further studying the importance of simulating ToM in human-robot communication.

1. Introduction

With advances in Artificially Intelligent (AI) agents and machine comprehension, social robots can be enhanced by intelligent conversational systems to provide fluid and natural conversations to users in different settings (Dino et al., 2019). To translate human behavior into computational algorithms through agent-based modeling, the majority of emerged artificial agents attempt to apply cognitive models to develop human-inspired intelligence. These models mainly rely on neural network based computation, such as machine learning, deep learning, or model-based reinforcement learning. These methods employ nonlinear continuous functions to regulate data and identify patterns. In this process, the system provides little transparency into the internal process of understanding how these machines make these decisions (Calder et al., 2018). Therefore, explaining why AI agents exhibit certain behaviors is always a challenge (Rai, 2020). From a user's viewpoint, the robot's behaviors or feedback may be unexpected. The lack of transparency in these models can impede users' trust since they are unable to understand or predict the robot's behavior. In other words, when users do not understand the cause or function of the robot's behaviors or decisions, they will lose trust in the robot (Miller, 2019). Trust is a significant and desirable characteristic of human-robot interactions. The lack of trust may further influence the user's acceptance of the robot's input, and reduce the efficacy of developing intuitive interaction (Song & Luximon, 2020).

Theory of mind (ToM) is a psychological trait that relates to developing interpretability in human communication. It refers to the ability of an individual to attribute mental states to others (e.g., beliefs, goals and desires) (Foss & Stea, 2014). Several studies have reported the close relationship between ToM and referential communication skills in cooperative tasks (Maridaki-Kassotaki & Antonopoulou, 2011; Paal & Bereczkei, 2007). Referential communication skills refer to the capacity to verbally transmit the representation of an object,

event or idea to a conversational partner to constitute the benchmark of a message (Liu et al., 2022). Referential communication tasks (RCTs) are used to evaluate the referential communication skills. A traditional RCT is usually conducted with two interlocutors who will act as speaker and listener in turns. Both the speaker and listener need to achieve a collaborative joint goal that ensures that their partner identifies the target referent. During this process, the speaker and listener must establish a shared understanding of the intended referent through verbal communication. Therefore, both interlocutors need to model their partners' viewpoint and adjust their own language accordingly to help each other identify the target referent. Because the task requires understanding the other's point of view, it necessarily involves ToM.

According to previous studies regarding RCTs, ToM skills are related to the communicative behaviors of requesting clarification and giving related information which refers to a communicative strategy called joint review (JR) (Sidera et al., 2018). With the inspiration of the association between JR and ToM skills, we attempt to develop a near-human response system to reinforce the representation of ToM in a robot-mediated RCT by increasing the appearance of JR to further enhance users' trust. Based on the theoretical model of JR, the robot needs to understand the user's description and provide explainable requests for clarifications or convincing information. Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is developed to focus on learning contextual relations between words, which has been validated as having significant performance accuracy in semantic awareness (Zhang et al., 2020). In this study, we present a novel approach with the application of BERT to develop the proposed response system.

2. The Robot-mediated RCT Experiment

The robot-mediated RCT experiment was conducted by participants interacting with a humanoid robot, Pepper. Each participant went through two phases: a *sorting* phase, and a *testing* phase. During the sorting phase, 12 abstract images were shown on Pepper's tablet (see Fig. 1, left panel). The 12 images were created with multiple objective characteristics which can be described with different descriptors. The robot described 3 images out of 12 shown on the screen to the participant, and the participant was asked to tap on the described image accordingly. If the participant selected the wrong image, Pepper would describe the image with a longer description which included more details. If the participant still could not select the correct target image after three rounds of description, Pepper would move on to the next image. The purpose of the sorting phase was to guide the participant in understanding how to communicate with Pepper in the following testing phase. Participants in each group would identify the same three images in the same order. In the testing phase, Pepper would show four abstract images on the screen for each trial. One image was highlighted by a black box which defined it as the target image (Fig. 1, right panel). The participant would organize his/her language to verbally describe the target image to Pepper. All four images could not be easily named or identified with simple labels, but contained different features to be described. Therefore, it was natural to observe participants describing the target image with different words. For example, the target image shown in the right panel can be described as "*keychain*" or "*five circles connected with each other*". A designed AI-mediated agent (Liu et al., 2021) would analyze both the transcript from the participant and the four images shown on the screen using a multi-modal vision-and-language analysis model and output four probability scores regarding the possibility of each image that the agent believed was the target image. Once the score

for one image was significantly higher than the others, the agent determined that confidence was high enough to select the image with the highest score as the target and say "*I think I found it. Let us move on to the next image*". It would then continue to the next set of four images. If none of the images has a score that is significantly higher than the other images, the robot would ask for more details from the participants by saying "*Could you give me more details?*" The participants would normally have to change their language to describe the same target image based on their predictions of the robot's understanding. Participants perceive the robot as having some level of intelligence since it can understand joint representations between their descriptions and the image referents. The agent analyzed all words the participant used for the current target image as input. Each time that the participant gave the robot a description was counted as one **round**. If Pepper still could not figure out the target image after three rounds, the system would automatically move on to the next trial.

Fig. 1. An example of the sorting phase (left panel) and the testing phase (right panel)

The testing phase had 24 trials. Among the 24 trials, the three abstract images used in the sorting phase were included as target images. All the images shown in the testing phase were presented in a pre-determined order. Therefore, all participants in the same group saw the same sequence of 24 trials. For each trial, a participant may have 1, 2, or 3 rounds. Participants' speech in each round during each trial was audio-recorded.

3. Proposed Method

In this study, we developed a system to generate natural language responses in a robot-mediated RCT. Inspired by ToM, we created a dialog system that allows robots to effectively communicate and engage with human users. The overall workflow of the designed dialog system is demonstrated in Fig. 2.

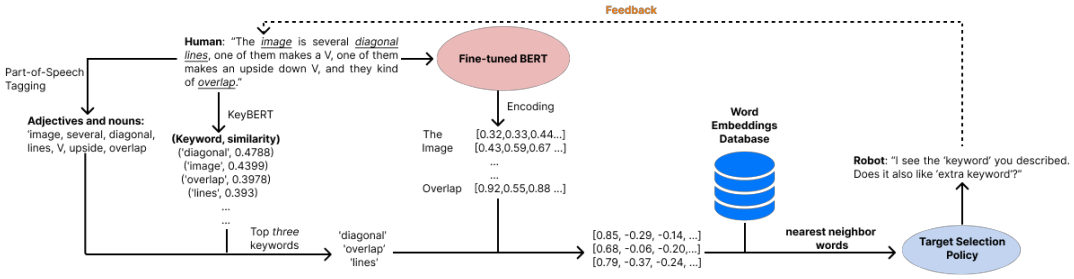

Fig. 2. Human-Robot Dialog System Workflow

The dialogue management module analyzes the input from the users and produces the corresponding response. The transcript is firstly analyzed via the designed keyword extraction approach to identify keywords which contain core semantics and are significant to the sentence. The word embeddings of the three most significant keywords are encoded by BERT and compared with every word in the word-embedding corpus using cosine similarity to find the three words (**extra keywords**) sharing the most relevant semantic meaning. The list serves as a reference for providing feedback includes "information the robot has understood" and "information the robot requests". For example, I see the "**keyword**" you described. Does it look like "**extra keyword**"?

The dialog architecture contains two key components: (1) perspective taking (Section 3.1) and (2) representation construction (Section 3.2). The architecture is trained with the dataset collected in a robot-mediated RCT (Section 4.1).

3.1. Perspective Taking

To develop a dialog system that can provide near-human response, the agent needs to understand the users' conceptual interpretation of the image. Keywords represent the specific semantics directly as a minimum knowledge unit. Moreover, they are valid and timely for tracking the information exchange among knowledge barriers (Xu et al., 2018). Therefore, keyword attentions are widely applied in conversation understanding (Mou et al., 2016). In the current study, we aim to extract context-based keywords as the representation of semantics to prove robots' understanding level of users.

KeyBERT is a state-of-the-art keyword extraction method that uses BERT embeddings to extract keywords that are the most representative of the underlying text document (Grootendorst, 2020). As shown in Fig. 2, the user's transcript is firstly analyzed via KeyBERT to identify keywords that contain more semantic meanings than other words and connote the main idea of the sentence. After obtaining the document-level representation (i.e., the sentence embedding) from BERT, KeyBERT extracts word embeddings and calculates their cosine similarity with the sentence. A term with the highest value is considered the one best representing the subject of the input.

According to Kennedy's research (Kennedy, 2007), vagueness complicates the processing of linguistic reasoning. To extend linguistic reasoning to vague description, we manually operationalized a contextually-determined threshold. Due to the experimental settings in RCT, participants would only provide a description of the target image. The context would naturally include the semantic representations of the target image. Since all the target images are black & white abstract images, the shape (e.g., circle) and object words (e.g., keychain) contain more conspicuous semantic features than other tokens in the sentence. As nouns and adjectives contain the most information about shape and object, we filtered the transcript input with part-of-speech (POS) tagging. Only the top three nouns and adjectives with the highest significance were selected as keywords.

3.2. Representation Construction

Simply demonstrating that the robot can identify the correct target image is not sufficient to simulate ToM. Based on the concept of ToM, the robot needs to attribute the mental states to the user (Foss & Stea, 2014). Referring it in the RCT, the robot is required to construct a representation of users' description from their point of view. In other words, the robot needs to provide extra information relevant to user's description. For example, the user describes: "*It is a keychain*". With a JR strategy, the robot is expected to speak like: "*Does the keychain* (related to user's description) *have a circle shape* (extra information) ?" To allow the robot give extra information, we generate a corpus containing semantic feature of transcripts collected before. The purpose of the corpus is to compare with the extracted keywords and select the word from the corpus with the most similar semantic representation as **extra keyword**.

Word embedding numerically captures the semantic relations between words. Words with similar meanings are proximate in the embedding space (i.e., nearest neighbors) (Fu et al., 2014). The nearest neighbors of a word indicate the meaning of the word in the context. Alternatively, they collectively represent a form of knowledge. Therefore, such a corpus allows us to calculate and compare the semantic correlation between existing transcripts in the corpus and the user's input in the RCT. Due to the advantage of semantic awareness, as shown in Fig. 2, we fine-tuned and applied BERT to extract word embeddings from transcripts in the dataset.

In order to mimic the natural communication in RCT, transcripts of each round are used as inputs. To ensure the applicability of this corpus under different settings, the two sets of transcripts were combined and collapsed as one dataset. By manually selecting words that are objects and shapes from the transcripts, we ensured that the word embeddings contained informative semantics about the target image. Only these words' embeddings are saved in the word-embedding corpus. The word embeddings serve as vector representations of their contextual semantics.

4. System Validation and Results

We applied two practiced BERT fine-tuning approaches: 1) BERT-ITPT-FIT (within-task-pre-trained and then fine-tuned) and 2) BERT-FIT (direct fine-tuned) since further pre-training could improves performance in downstream natural language processing tasks (Sun et al., 2019). We evaluated the designed dialogue system from two different perspectives: (1) transcript

classification performance, and (2) dialog simulation performance. Performance metrics were calculated based on the prediction in the last training epoch.

4.1. Dataset

A dataset with 96 young adults' speech transcripts during a robot-mediated RCT was applied in this study. All participants were native speakers of English and recruited from a large engineering course offered at a large state university in the southeast US. The participants were randomly and evenly divided into two groups. In each group, a robot-mediated RCT was conducted with the same protocol but different image sets (see Section 2 for details). The study protocol was approved by the Institutional Review Board (IRB) of the University of Tennessee, Knoxville (UTK IRB-21-06631-XM).

4.2. Transcript Classification Performance

We validated the fine-tuned BERT models' abilities to extract words' semantic representation. Since the text-classification results is correlated with BERT model's semantic awareness, the BERT-based classifier's performance was measured, considered as the performance of its semantic representation.

BERT-ITPT-FIT and BERT-FIT were compared in terms of their ultimate performance in the downstream 48-class-transcript classification task, regarding the 48 trials conducted in two sets of RCT. A 10-fold cross-validation was conducted and the classification accuracy, precision, recall,

and F1 score were used to compare the performance. We also tested four models trained on different subsets of the dataset (referred to as subset-models) and performed 10-fold cross validation for each model to mitigate the effect of randomness.

As summarized in Table 1, both BERT-ITPT-FIT and BERT-FIT achieved an average accuracy over 84% in the 10-fold cross-validation. The average precision, recall, and F1 scores also surpassed .84 and validated the model's performance in classifying the transcripts. There was no significant effect for whether conducting within-task pre-training on the mean classification accuracy ($t(18) = .29$, $p > .05$), macro precision ($t(18) = .20$, $p > .05$), macro recall ($t(18) = .23$, $p > .05$), or macro F1 score ($t(18) = .38$, $p > .05$), despite that BERT-ITPT-FIT attained slightly higher average scores across all metrics except for the subset-model accuracy than BERT-FIT. Further experiments on the four subset-models revealed no significant difference ($t(78) = .55$, $p > .05$) in classification accuracy between the models trained by different methods.

Model	Accuracy	Precision	Recall	F1	Subset-model Accuracy
BERT-FIT	. 8468 (.0276)	.8433 (.0275)	.8650 (.0251)	.8452 (.0277)	**.8186** (.0310)
BERT-ITPT-FIT	**.8503** (.0256)	**.8458** (.0279)	**.8675** (.0238)	**.8497** (.0250)	.8149 (.0293)

Table 1. Transcript Classification Performance, *M* (*SD*)

4.3. Dialog Simulation Performance

The dialog system was also evaluated based on how well the system can determine the additional keywords. The testing dataset was applied as transcript inputs into the designed dialog system. If one of the three relevant words determined by the system exists in the transcripts from the training

dataset which described the same target image, it would count as a match. Otherwise, it was not a match. The proportion of simulation transcripts that contain at least one match (referred to as match ratio hereafter) was used as the criteria of comparison.

Due to the assumption of object and shape words, we plan to validate the system by calculating the values when the extracted keywords with and without any shape and object words (refers to normal and unexpected situation, respectively). The shape and object list were created by manual selection from transcripts in the dataset for each training set to maintain the consistency in the training-simulation data partition. Each model and shape/object list being used by that model was trained and generated from a subset of the whole data so that the simulation transcripts were not accessible in the training process and served as unseen sentences merely for validation.

When BERT is fine-tuned for a downstream task, token representation is one of the salient factors affecting its performance since different layers of the BERT model output different semantic features. We tested two approaches to represent every token: (1) only using the output features (hidden state of the BERT encoder) from the last layer and (2) summing all the output features from the last four layers.

Fig. 3 gives a summary of evaluation results regarding the dialog simulation performance. Overall, although models fine-tuned after within-task-pre-training (BERT-ITPT-FIT) yielded slightly higher mean match ratio and reliability than models directly fine-tuned (BERT-FIT) for text classification, there was no significant effect ($t(30) = 1.01$, $p > .05$) for their average performance in normal and worst situations altogether. Nor for the differences between BERT-ITPT-FIT and

FIT with the sum of the last four layers to represent tokens ($t(14) = .35386$, $p >. 05$) and with the last layer ($t(14) = 1.36575$, $p > .05$). It is therefore arguable that within-task-pre-training contributed limited classification capacity in this particular task. One explanation could be the small volume of the training data which made it inefficient transferring the BERT language model to this specific domain.

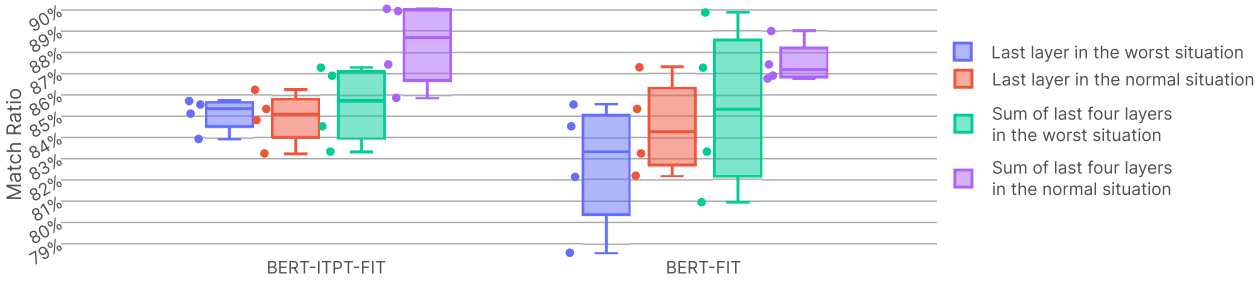

Fig. 3. Simulation results for the normal and worst situations

Despite the limited performance improvement of using within-task pre-training in our task, we observed significant differences in match ratios between outcomes with different token representations. Experiments using the sum of the last four layers' features produced significantly higher mean match ratio ($t(30) = 2.87$, $p = .0075$) than only using the last layer's features in normal and worst situations altogether. The BERT-ITPT-FIT model with sum of the last four layers' output features as the token representation attained the highest match ratio of 90.05% (matched 172/189 transcripts). The significant difference also holds when comparing the mean match ratio for each model training approach: BERT-ITPT-FIT ($t(14) = 2.13$, $p = .0518$) and BERT-FIT ($t(14) = 2.01$, $p = .0636$). Our simulation results are consistent with the conclusion that the sum of the last four layers capture richer semantic meanings in a variety of levels than solely the last layer (Devlin et al., 2018). Based on previous and our simulation results, a natural hypothesis would be that representing the tokens by concatenating the last four layers' output features could further improve

the match ratio in simulation and the overall performance in field experiments, which could be tested for the following work.

The performance of the dialog system in simulation was evaluated by metrics of the transcript classification and ability to find related descriptions as response. We found similar text classification metrics between the BERT-ITPT-FIT and BERT-FIT model. It demonstrated that the model could identify the label of the described image by encoding and classifying the text features. The encoded features are therefore distributed in a way that makes it suitable to find relevant descriptions through vector similarities. The dialog simulation results further evidenced the system's capacity to find relevant and coherent words to form the response. Moreover, we found that representation of the tokens had an effect on performance in the dialog simulation.

To summarize, our simulation preliminarily confirmed the validity of the proposed dialog system in finding relevant words to facilitate a jointly reviewed conversation in RCT. The effect of model-training approach and token representation was analyzed. More training data and alternative token representation methods will be explored in the future study.

5. Discussion and Conclusion

We developed a robot dialog system for RCT based on the mechanism of ToM applied in daily human-human communication. The aim of the proposed dialog system is to enhance the user's understanding of robot's intention, and further improve users' trust towards the robot. Regarding the results from two validation tests, the designed system contains the capability of semantic

awareness to understand the concepts represented in users' descriptions and the ability to determine extra keywords necessary for clarification. Therefore, we believe the designed system bears an acceptable performance to conduct the proposed dialog and can be a research tool for the future studies in human-robot communication. Further field study is needed to test ecological validity of the dialog system to understand how it impacts the trust level of users in real-life conditions and interactions between AI agents and humans.

Acknowledgement

This project was sponsored by the National Science Foundation through the Research Experience for Undergraduates (REU) award no. 2020534 with additional support from the National Institute of Computational Sciences and Innovative Computing Laboratory at the University of Tennessee, Knoxville.

References

Calder, M., Craig, C., Culley, D., De Cani, R., Donnelly, C. A., Douglas, R., Edmonds, B., Gascoigne, J., Gilbert, N., Hargrove, C., & others. (2018). Computational modelling for decision-making: Where, why, what, who and how. *Royal Society Open Science*, *5*(6), 172096.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv. https://doi.org/10.48550/ARXIV.1810.04805

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dino, F., Zandie, R., Abdollahi, H., Schoeder, S., & Mahoor, M. H. (2019). Delivering cognitive behavioral therapy using a conversational social robot. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2089–2095.

Foss, N., & Stea, D. (2014). Putting a realistic theory of mind into agency theory: Implications for reward design and management in principal-agent relations. *European Management Review*, *11*(1), 101–116.

Fu, R., Guo, J., Qin, B., Che, W., Wang, H., & Liu, T. (2014). Learning semantic hierarchies via word embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1199–1209.

Grootendorst, M. (2020). *KeyBERT: Minimal keyword extraction with BERT.* (v0.3.0). Zenodo. https://doi.org/10.5281/zenodo.4461265

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*(1), 1–45.

Liu, Z., Collier, P. N., Wang, C., Paek, E. J., Yoon, S. O., Casenhiser, D., & Zhao, X. (2021). A demonstration of human-robot communication based on multiskilled language-image

analysis. *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 126–127.

Liu, Z., Paek, E. J., Yoon, S. O., Casenhiser, D., Zhou, W., & Zhao, X. (2022). Detecting Alzheimer's Disease Using Natural Language Processing of Referential Communication Task Transcripts. *Journal of Alzheimer's Disease*, *86*(3), 1385–1398. https://doi.org/10.3233/JAD-215137

Maridaki-Kassotaki, K., & Antonopoulou, K. (2011). Examination of the relationship between false-belief understanding and referential communication skills. *European Journal of Psychology of Education*, *26*(1), 75–84.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Mou, L., Song, Y., Yan, R., Li, G., Zhang, L., & Jin, Z. (2016). Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *ArXiv Preprint ArXiv:1607.00970*.

Paal, T., & Bereczkei, T. (2007). Adult theory of mind, cooperation, Machiavellianism: The effect of mindreading on social relations. *Personality and Individual Differences*, *43*(3), 541–551.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141.

Sidera, F., Perpiñà, G., Serrano, J., & Rostan, C. (2018). Why is theory of mind important for referential communication? *Current Psychology*, *37*(1), 82–97.

Song, Y., & Luximon, Y. (2020). Trust in AI agent: A systematic review of facial anthropomorphic trustworthiness for social robot design. *Sensors*, *20*(18), 5087.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification? *China National Conference on Chinese Computational Linguistics*, 194–206.

Xu, J., Bu, Y., Ding, Y., Yang, S., Zhang, H., Yu, C., & Sun, L. (2018). Understanding the formation of interdisciplinary research from the perspective of keyword evolution: A case study on joint attention. *Scientometrics*, *117*(2), 973–995.

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., & Zhou, X. (2020). Semantics-aware BERT for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 9628–9635.