Robot-mediated Referential Communication: To Improve Trust in Human–robot Interaction

Students:
Yigang Qin, Huiqi Zou

Mentors:
Dr Xiaopeng Zhao, Ziming Liu,
Dr Kwai Wong
Trust AI robot?
A prosthetic arm displayed at an exhibit about science and technology changing perceptions of humanity.
The Black Box Society: The Secret Algorithms That Control Money and Information
Do we know the thinking of an AI model or a robot?
Socially Assistive Robot (SAR)

Fairness

Accountability

Transparency

Ethics
Improve human’s trust in robot-mediated referential communication task
Table of Contents

1. Concept Map
   - Theory of Mind
   - Referential Communication
   - Joint Review

2. Experiment
   - Experiment Setting

3. System Architecture
   - Data and System Architecture

4. Validation
   - Metrics and Results
1 Concept Map
Theory of Mind, Referential Communication, and Joint Review,
Theory of Mind

The basic cognitive and social characteristic that enables us to make conjectures about others’ minds through observable, or latent behavioral, and verbal cues.
Philosophical root: Philosophy of Mind

René Descartes. *Meditations on First Philosophy*. 1641

The Nature of Human Mind
The ability to take someone else’s perspective

- *empathy*: the ability to understand and share the feelings of others
Theory of Mind
Theory of Mind

- Recognizing other’s feelings
- Thinking about consequences of actions
- Recognizing that someone else may think or feel differently than you do

Children’s theory of mind in development
Theory of Mind

Computational Theory of Mind

Level -1
We act on an impersonal environment

Level 0
What type is our partner?

Level 1
What type do they think we are?

Level 2
What type do they think, we think, they are?
Referential Communication

Communicative action of referring to ... something
Referential Communication
The most used communicative strategy in Referential Communication Task is Joint Review which is closely related to the Theory of Mind.
Knowledge Pertaining to the Materials Used in the Referential Communication Task

Director's Knowledge

Matcher's Knowledge

Director's Assumptions About the Matcher's Knowledge

A + D: Director's Unique Knowledge
C + F: Matcher's Unique Knowledge
D + E: Director Assumed Shared Knowledge
B + E: Actual Shared Knowledge
D: Overestimated Shared Knowledge
E: Correctly Assumed Shared Knowledge
B: Underestimated Shared Knowledge
F: Correctly Assumed Matcher's Knowledge
G: Overestimated Matcher's Unique Knowledge
H: Shared Ignorance

Referential Communication

Joint Review

Theory of Mind

Understanding and Trust
Experiment

Referential Communication Task
Sorting Phase

Purpose: Guide the participant in understanding how to communicate with the robot
Testing Phase

An example that the robot provides extra information relevant to participant’s description.

It is a keychain.

Is the keychain *(related to user’s description)* with a circle shape *(extra information)*?
3 System Architecture
Data and System Architecture
A dialog system that can provide near-human response

KeyBERT

The minimum knowledge unit to represent the specific semantics

Bidirectional Encoder Representations from Transformers

Feedback

Target Selection Policy

Part-of-Speech Tagging

Adjectives and nouns: 'image, several, diagonal, lines, V, upside, overlap'

(Keyword, similarity)

'diagonal', 0.4788
'overlap', 0.3978
'lines', 0.393

Top three keywords

'diagonal' 'overlap' 'lines'

Encoding

The Image

[0.32, 0.33, 0.44, ...]

Overlap

[0.92, 0.55, 0.88, ...]

Word Embeddings Database

nearest neighbor words

Robot: "I see the 'keyword' you described. Does it also like 'extra keyword'?"

Human: "The image is several diagonal lines, one of them makes a V, one of them makes an upside down V, and they kind of overlap."
Word Embedding

numerically captures the semantic relations between words
The minimum knowledge unit to represent the specific semantics.
Validation

Metrics and Results
## Factor

<table>
<thead>
<tr>
<th>Token representation</th>
<th>BERT training approach</th>
<th>Information saturation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use the output features from the last layer</td>
<td>Within-task-pre-trained and then fine-tuned (BERT-ITPT-FIT)</td>
<td>Normal situation</td>
</tr>
<tr>
<td>Sum all the output features from the last four layers</td>
<td>Direct fine-tuned (BERT-FIT)</td>
<td>Worst situation (Shape and object words excluded)</td>
</tr>
</tbody>
</table>
## Transcript Classification

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Subset-model Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-FIT</td>
<td>.846761 (.027618)</td>
<td>.843257 (.027536)</td>
<td>.864997 (.025090)</td>
<td>.845208 (.027731)</td>
<td>.818619 (.030962)</td>
</tr>
<tr>
<td>BERT-ITPT-FIT</td>
<td>.850260 (.025637)</td>
<td>.845776 (.027907)</td>
<td>.867514 (.023780)</td>
<td>.849688 (.024957)</td>
<td>.814890 (.029276)</td>
</tr>
</tbody>
</table>

10-fold Cross-validation metrics M (SD) on 48-class transcript classification
Dialog Simulation

Match: one of the three extracted words is in the transcripts from the training dataset describing the same target image

Simulation results for the normal and worst situations
System Features

Understand the users’ descriptions

Extract keywords for clarification

Enhance users’ understanding on robot’s intention

Improve users’ trusts towards the robot
References


Thanks!
Awesome Questions?