Teaching a Small Robot how to Understand Spoken Language

BOISE STATE UNIVERSITY
COLLEGE OF ENGINEERING
Department of Computer Science

Casey Kennington, PhD and Sarah Plane
Boise State University

SPEECH, LANGUAGE & INTERACTIVE MACHINES

Language Acquisition in Robots

Current systems do not reflect the natural, interactive process of child language acquisition.

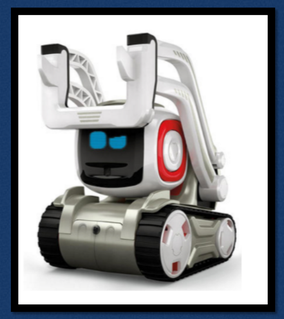
- Systems are trained on large amounts of textual data
- The system is evaluated using speech & ASR, often unsuccessfully

Face-to-face interaction is crucial for grounding, which is key when establishing meaning.

- **Symbol grounding**: language is connected to aspects of the object (sight, smell, touch, etc.).
- **Conversational grounding**: aspects of events are recorded for later recall

Solution: Use a grounded semantic model to teach a robot about meaning through online learning.

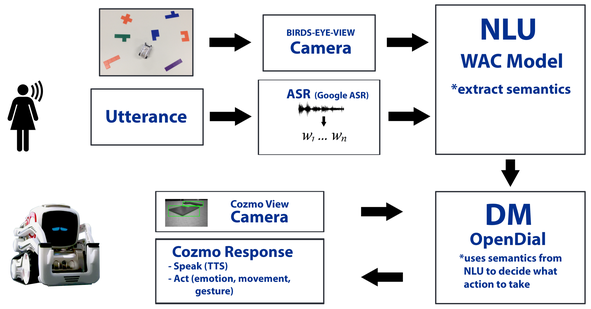
- **An interactive, incremental SDS using WAC**

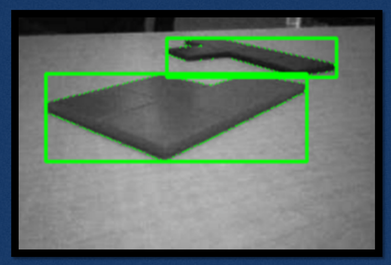


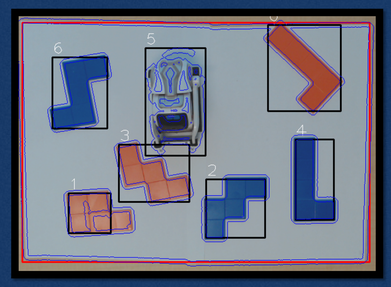
Why Cozmo?

- Cozmo is small and portable
- SDK & support community
- Can lift, push, and point
- Text to speech synthesizer
- Black and white camera
- Affordable
- Cozmo is adorable/likeable

Spoken Dialog System







Initial Testing

Training: 300 episodes to train the WAC model

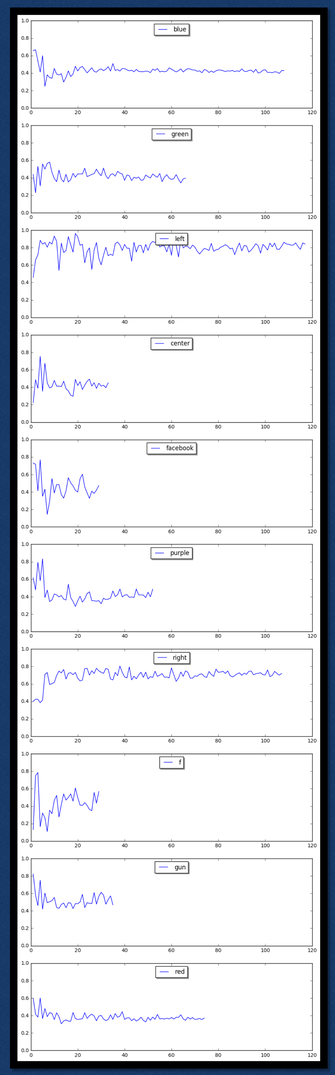
Evaluation:

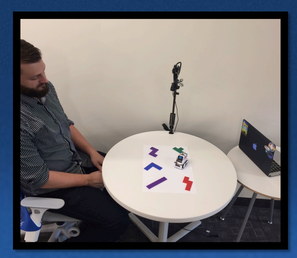
- 89 episodes (untrained)
- 29 WAC classifiers
- Object features were fed to the WAC classifiers giving a probability.
- Compared to actual referring expression
- Convert average ranks to an MMR to get a score between 0 – 1

Results:

- Right & Left (MMR ~ 0.8)
- Shapes (MMR ~ 0.5)
- Color (MMR under 0.5)

Due to the low MMR for color we added a color camera with a bird-eye perspective to help with color detection.





Words-As-Classifiers (WAC) Model

Model of Word Meaning:

1. Single object via visual properties: takes visual features & returns a probability for the word to match the features.
2. Relation between two objects: takes a vector of features for a set of objects to train on Euclidean distance, vertical and horizontal differences, and spatial relationships.

Trained using the referring expression and the visual scene, resulting in a set of classifiers. When evaluated, it predicts the relationship between a word and an object.

Composition & Selection:

Distributes predictions over all candidate objects for an entire utterance.

1. Simple References: only contains properties for the referent object (*the red cross*)
2. Relational References: has properties for both a target and a landmark object (*the red cross to the left of the purple T*)

Evaluating with the WAC Model:

- Rank the probabilities for word match
- High Mean Reciprocal Rank (MMR)
- Compose & Select for DM

Benefits of WAC Model:

- Transparency
- Modularized
- Amendable and Scalable

The Experiment

Each participant will be asked to complete two tasks (given in a randomized order):

Task 1

Interactions will begin with simple reference resolution tasks (e.g. the participant will teach the robot what a new shape is: *the red cross on the left*) and will build to an interaction task in which Cozmo locates ("fetches") a specific requested object while navigating through a variety of objects (e.g., *Cozmo, find the purple P*). This models a more passive learning style, which is representative of the current systems for natural language processing models.

Task 2

Interactions will begin with the task in which the robot locates ("fetches") a specific requested object while navigating through a variety of objects (e.g., *Cozmo, find the purple P*). Cozmo will have to guess and check, with a more adaptive, active learning style, which is representative of how children acquire their first language. This task requires the robot to learn online through interaction with the participant.

Follow-up surveys will be conducted after each task to help us compare the participant's response to each learning style.

Each session will log:

- A png snapshot of the setting (as seen by Cozmo, B&W)
- A png snapshot of the object (cropped using Cozmo's object detection)
- A png snapshot of the birds-eye view
- A png snapshot of the object (cropped using the birdseye object detection)
- A Pickle (a compilation of a series of objects)
- Speech recorded from the participant & audio recorded from Cozmo
- Visual recording of the entire session captured by an external camera

HERC Fellowship funded by the Institute for STEM and Diversity Initiative at Boise State University

BOISE STATE UNIVERSITY

Contact: sarahplane@u.boisestate.edu