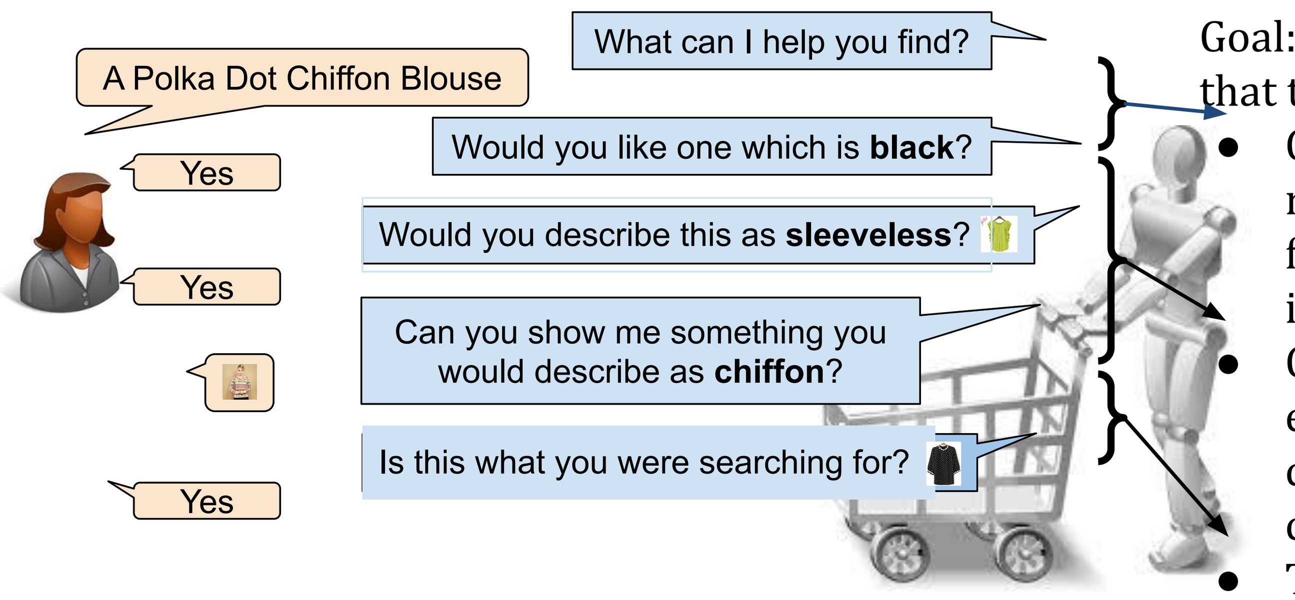
Dialog Policy Learning for Joint Clarification and Active Learning Queries

Aishwarya Padmakumar[1] **Raymond J. Mooney**[2]

[1]Amazon Alexa AI [2]University of Texas at Austin

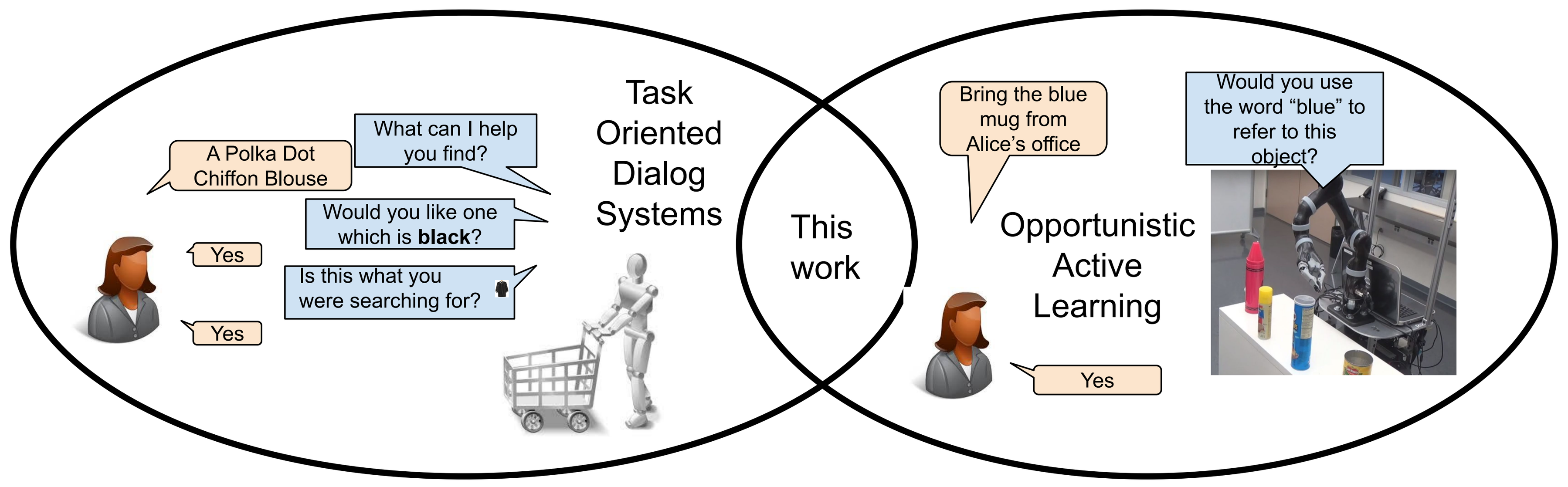
Dialog System Learning New Concepts



Goal: Learn a dialog policy that trades off:

- Clarification to get necessary information for current user interaction
- Collecting labelled examples to build classifiers for new concepts
- Task Completion

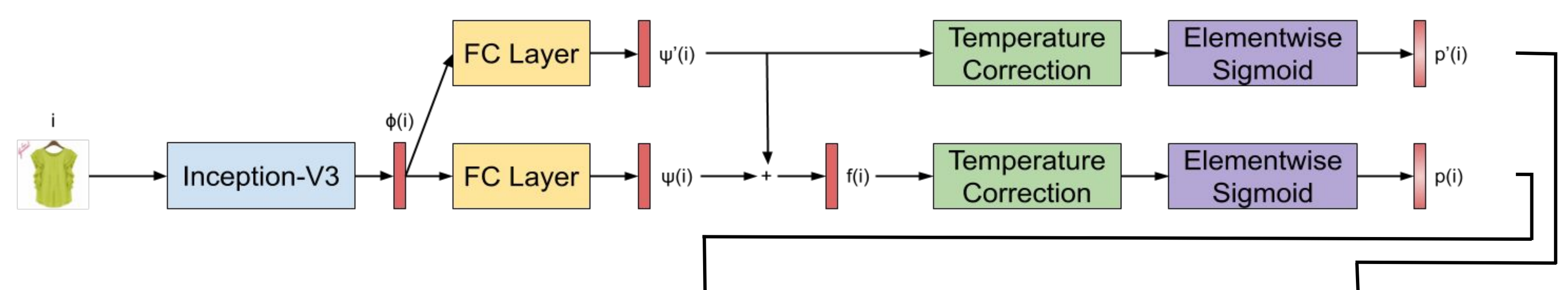
We augment task oriented dialog systems with opportunistic active learning queries to enable the system to generalize to new concepts not seen at training time.
For a dialog system in the shopping domain, this means automatically adapting to an expanded inventory with new product categories without collecting training dialogs on these categories.



Dialog Simulation

- We simulate dialogs using the iMaterialist Fashion Attribute dataset which has images with product titles and 228 binary attributes labeled.
- Product titles provided as simulated user queries.
- System actions:
  - Clarification - Is selected attribute is positive for target object?
  - Label Query - Is selected attribute positive for selected training image?
  - Example Query - Request for a training image which is positive for selected attribute.
- System questions answered using binary attribute labels.
- Dialog ends when system guesses the item the user was searching for.

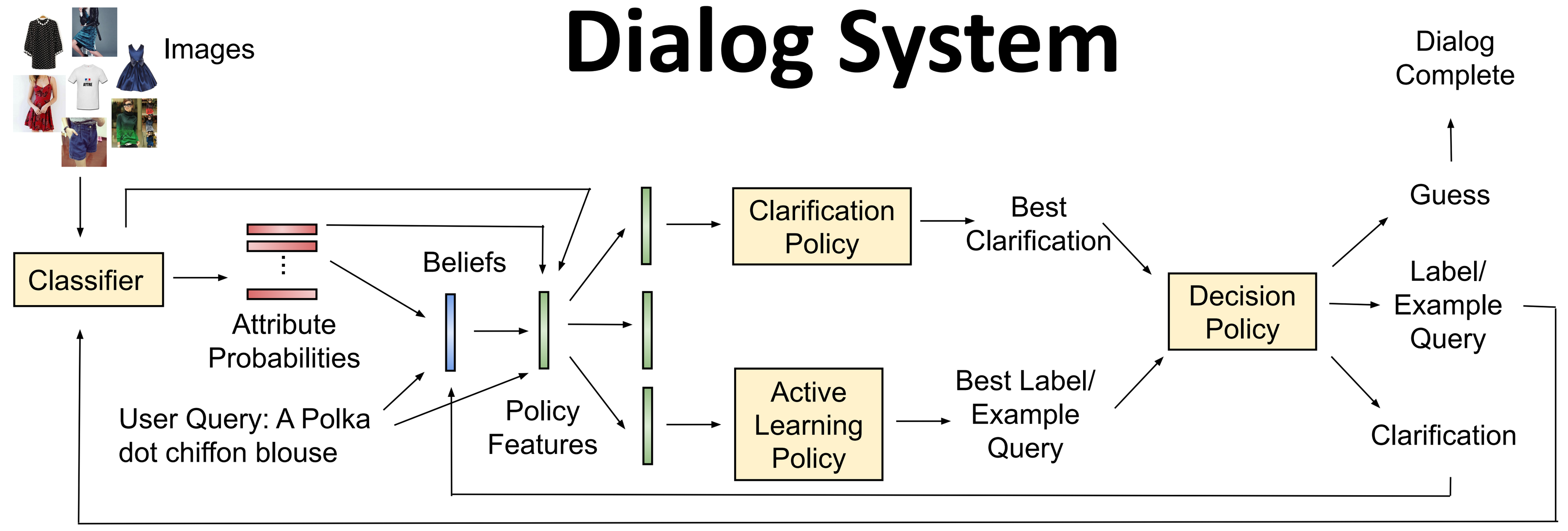
Visual Attribute Classifier



$$L = \boxed{(1-\lambda)\sum_i y_i \log p(i) + (1-y_i)\log(1-p(i))} + \boxed{\lambda \sum_i y_i \log p'(i)}$$

Two branch convolutional neural network to upweight positive examples

Dialog System



Policy Features:

- Clarification Policy Features - Metrics about current beliefs, information gain estimated from classifier probabilities
- Active Learning Policy Features - Margin, Fraction of previous uses and successes
- Decision Policy Features - Metrics about current beliefs, information gain, margin, dialog length

Static Baseline:

- Clarification: Choose query with maximum estimated information gain
- Active Learning: Uncertainty Sampling
- Decision Policy
  - Fixed dialog length
  - Dialog split equally between clarification and active learning
  - Heuristic checks to ensure usefulness of queries

Policy Learning:

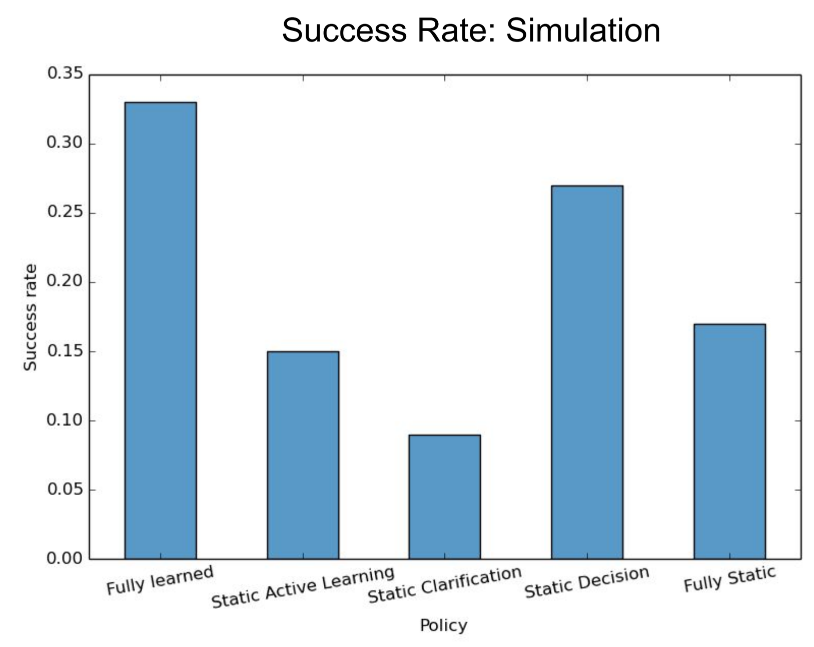
- Dialog modeled as MDP with featurized state-action pairs
- Reward: Large positive value for correct guess, large negative value for incorrect guess, small negative value for additional questions.
- RL algorithms: Q-learning and A3C

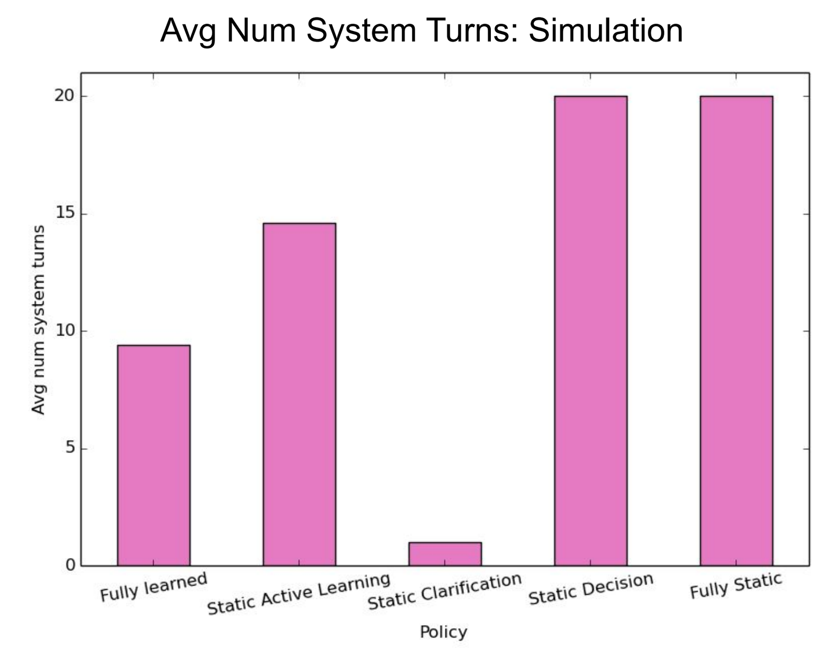
Experiment Phases:

- Classifier Initialization - Train classifier using paired images and labels
- Policy Initialization - Collect experience using the baseline to initialize the policy.
- Policy Training - Improve the policy from on-policy experience.
- Policy Testing - Policy weights are fixed, and we run a new set of interactions, reset classifiers to the state at the end of classifier initialization, over an independent test set containing novel attributes.

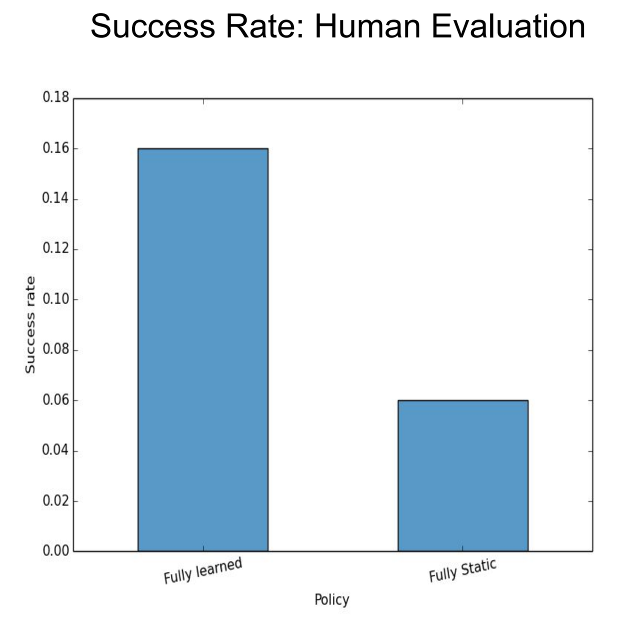
Results

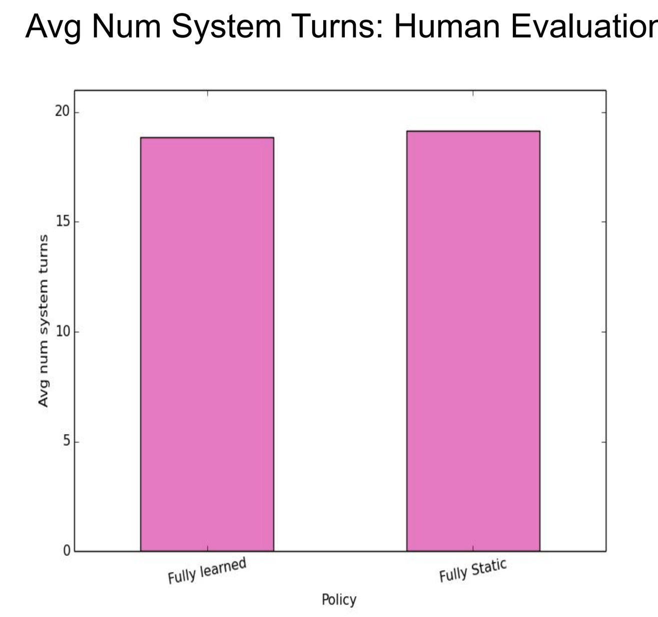
In simulation, our learned policy is more successful than a static baseline while using fewer dialog turns on average.





Neither the static nor the learned policies transfer well during human evaluation but the learned policy remains more successful than the static policy.





Utility of Clarifications

Only the combination of learned clarifications and learned active learning queries improve over direct retrieval.