An Embodied Incremental Bayesian Model of Cross-Situational Word Learning

Sepideh Sadeghi, Matthias Scheutz, and Evan Krause

Department of Computer Science, Tufts University

Medford, MA 02155, USA

Tufts UNIVERSITY

hrf lab

Word Learning in Ambiguous Contexts

Utterance: “Jack is eating the apple.”

Word Learning in Ambiguous Contexts

Utterance: “Jack is eating the apple.”

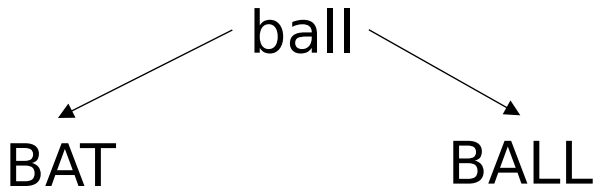
Word Learning in Ambiguous Contexts

Utterance: “Jack is eating the apple.”

Word Learning in Ambiguous Contexts

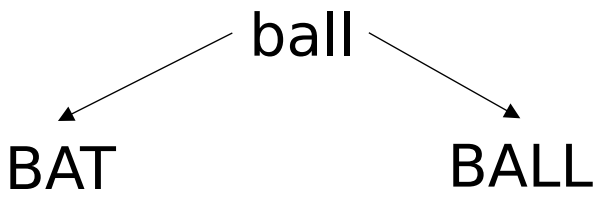
Utterance: “Jack is eating the apple.”

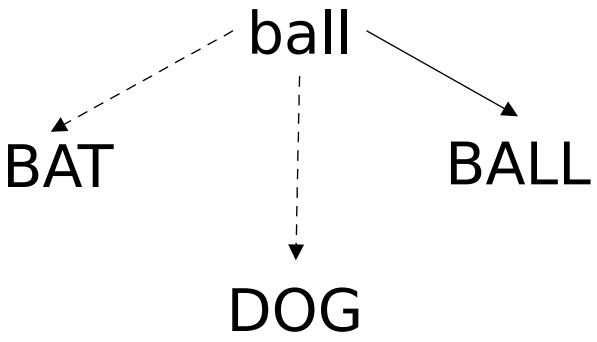
Mapping Words to Referents



Quine, W. V. O., *Word and object*, 1960

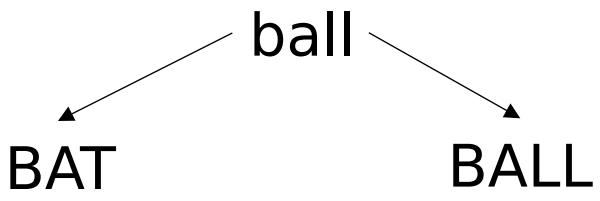
Mapping Words to Referents

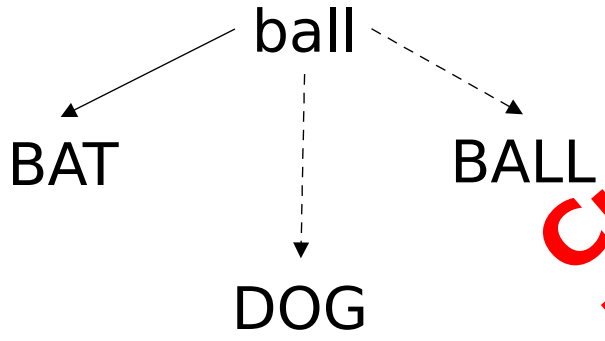




Quine, W. V. O., *Word and object*, 1960

Mapping Words to Referents

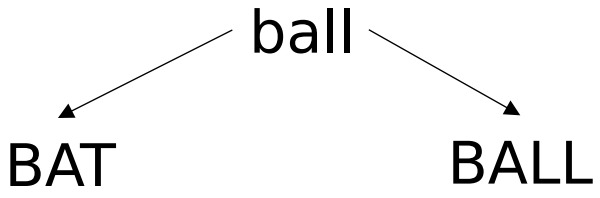




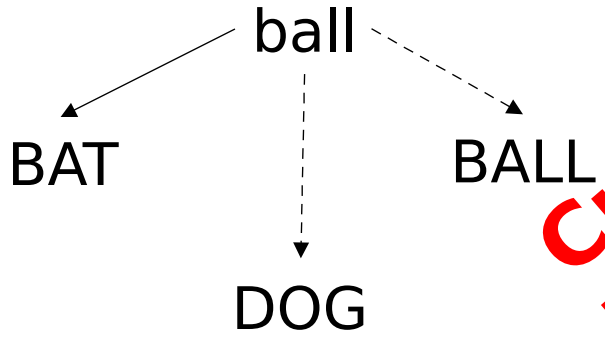
Cross-Situational Word Learning

Quine, W. V. O., *Word and object*, 1960

Mapping Words to Referents



Incremental

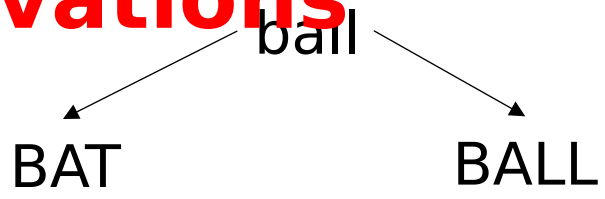


Cross-Situational Word Learning

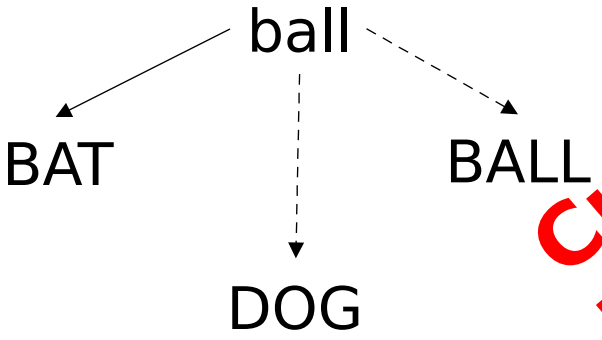
Quine, W. V. O., *Word and object*, 1960

Mapping Words to Referents

With limited memory of past observations



Incremental



Cross-Situational Word Learning

Quine, W. V. O., *Word and object*, 1960

Motivation

Past models of cross-situational word learning
(e.g., by Frank et al. 2009) benefit from **full perfect access**
to all learning situations and their statistical regularities
to arrive at the best word-object mapping hypothesis

However, it is **cognitively implausible**
for children to remember
all word learning situations they encounter

Moreover, the real world is noisy and we can thus not
assume perfect perceptual information

Agenda

(1) Learn words **incrementally from situation to situation** using an incremental Bayesian model of cross-situational word learning with limited access to past situations

(2) Learn words **in noisy contexts** with noise on both the auditory and the visual channels.

(3) Demonstrate the superior performance of our model compared to other baseline incremental models, especially under conditions of sensory noise in the speech and visual modalities

(4) Demonstrate that our model embedded in a cognitive robotic is capable of real-world incremental cross-situational word learning

Input Representation

Utterance: “Jack is eating the apple”

Scene: 

$$I_s \in P(O_s)$$

Situation=<utterance,scene>

Utterance = W_s= {jack, is, eating, the, apple}

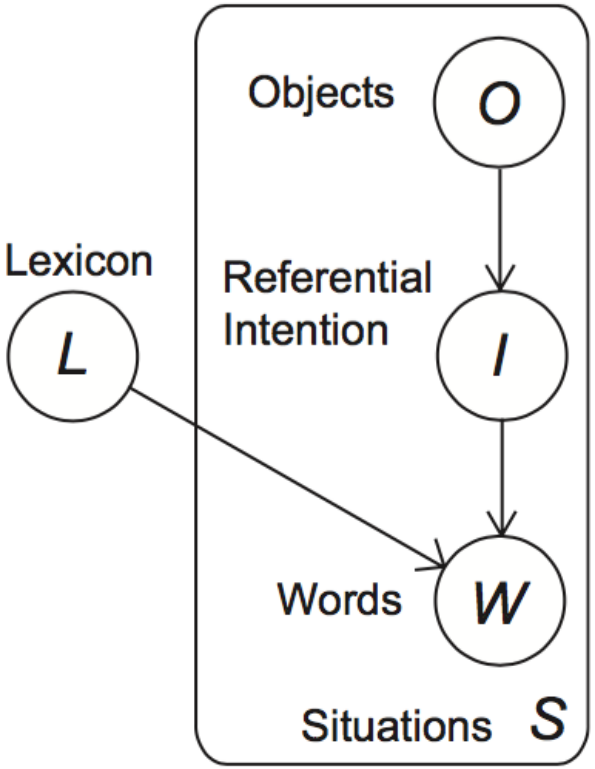
Scene = O_s= {JACK, SARAH, PLATE, APPLE, CHAIR, TABLE, NAPKIN, BISCUIT, MILK, GLASS}

I_S = {JACK, APPLE}

Model Design and Generative Process (based on Frank et al. 2009)

Scene = {JACK, SARAH, PLATE, APPLE, CHAIR, TABLE, NAPKIN, BISCUIT, MILK, GLASS}

I_S = {JACK, APPLE}



Utterance: “Jack is eating the apple”

scene:

Model Design and Generative Process

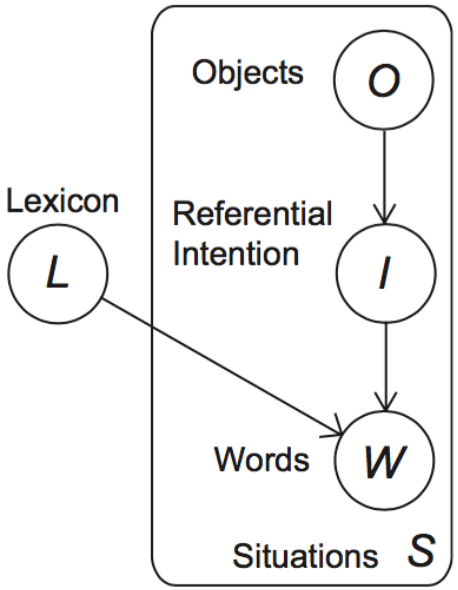
L = {Jack: JACK,apple: APPLE, sarah:SARAH}

I_S = {JACK, APPLE}

referential words

non-referential words

component in focus



scene:

Utterance: “Jack is eating the apple”

$P_R(w) = \gamma$

$P_{NR}(w) = 1\text{-}\gamma$

Model Design and Generative Process

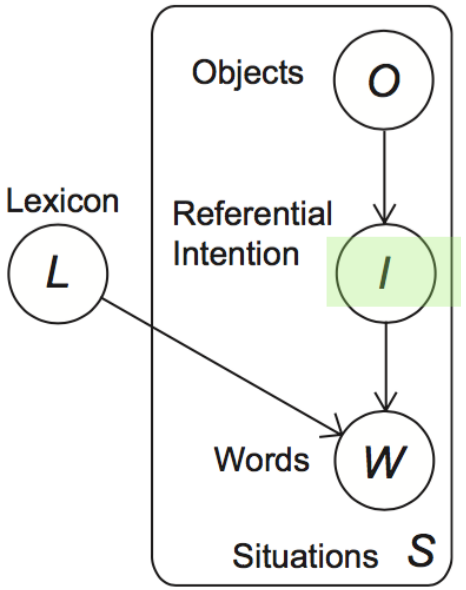
L = {Jack: JACK,apple: APPLE, sarah:SARAH}

I_S = {JACK, APPLE}

referential words

non-referential words

component in focus



scene:

Utterance: “Jack is eating the apple”

$P_R(w) = \gamma$

$P_{NR}(w) = 1\text{-}\gamma$

Model Design and Generative Process

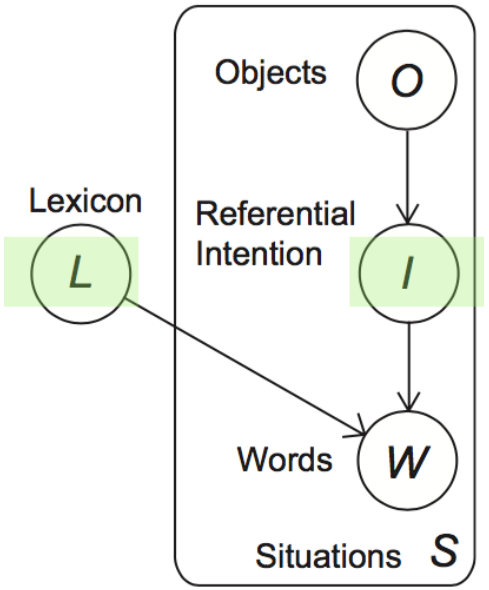
L = {Jack: JACK,apple: APPLE, sarah:SARAH}

I_S = {JACK, APPLE}

referential words

non-referential words

component in focus



scene:

Utterance: “Jack is eating the apple”

$P_R(w) = \gamma$

$P_{NR}(w) = 1\text{-}\gamma$

Model Design and Generative Process

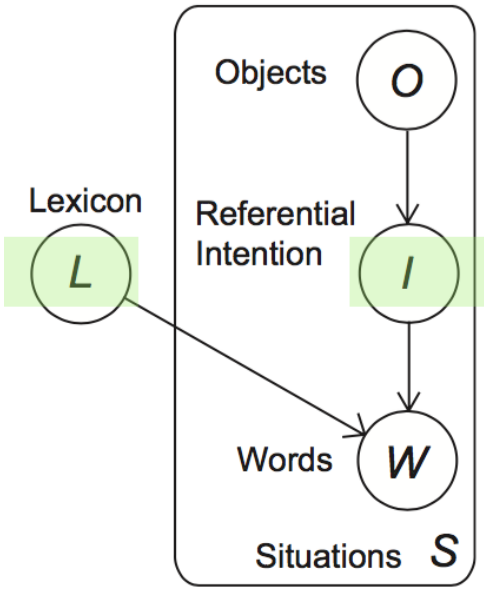
L = {Jack: JACK,apple: APPLE, sarah:SARAH}

I_S = {JACK, APPLE}

referential words

non-referential words

component in focus



scene:

Utterance: “Jack is eating the apple”

$P_R(w) = \gamma$

$P_{NR}(w) = 1\text{-}\gamma$

Model Design and Generative Process

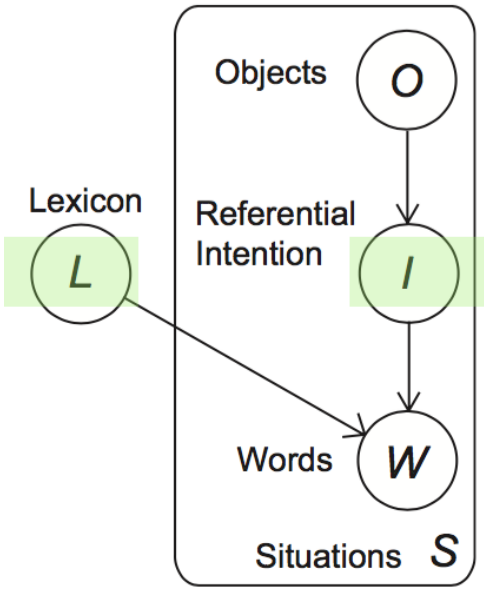
L = {Jack: JACK,apple: APPLE, sarah:SARAH}

I_S = {JACK, APPLE}

referential words

non-referential words

component in focus



scene:

Utterance: "Jack is eating the apple"

$P_R(w) = \gamma$

$P_{NR}(w) = 1\text{-}\gamma$

Model Design and Generative Process

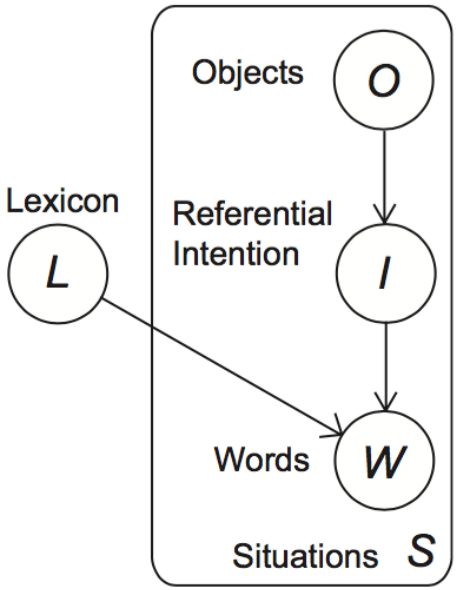
L = {Jack: JACK,apple: APPLE, sarah:SARAH}

I_s = {JACK, APPLE}

referential words

non-referential words

component in focus



scene:

Utterance: “Jack is eating the apple”

$P_R(w) = \gamma$

$P_{NR}(w) = 1\text{-}\gamma$

Model Design and Generative Process

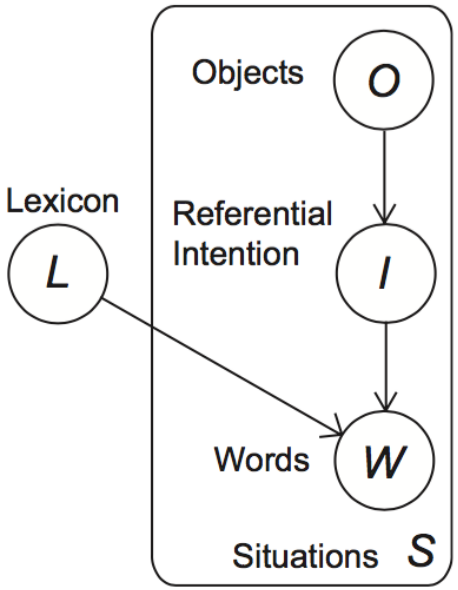
L = {Jack: JACK,apple: APPLE, sarah:SARAH}

I_s = {JACK, APPLE}

referential words

non-referential words

component in focus



scene:

Utterance: “Jack is eating the apple”

$P_R(w) = \gamma$

$P_{NR}(w) = 1\text{-}\gamma$

Model Design and Generative Process

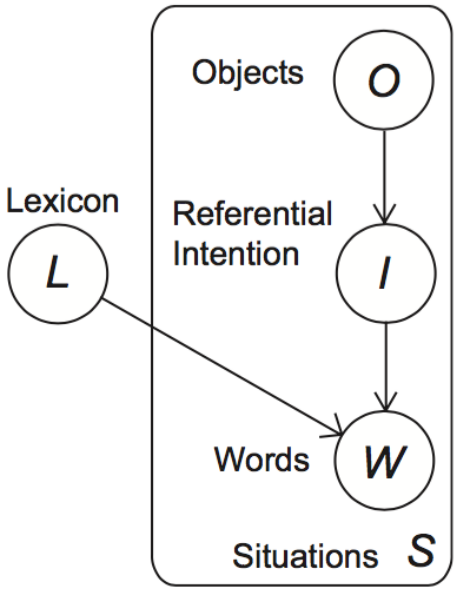
L = {Jack: JACK,apple: APPLE, sarah:SARAH}

I_S = {JACK, APPLE}

referential words

non-referential words

component in focus



scene:

Utterance: “Jack is eating the apple”

$P_R(w) = \gamma$

$P_{NR}(w) = 1-$

Model Design and Generative Process

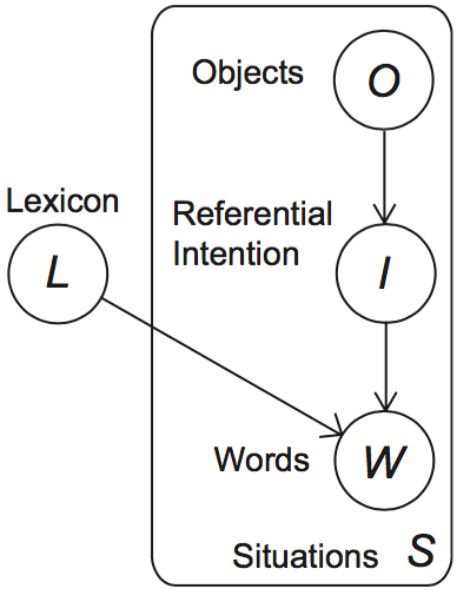
L = {Jack: JACK,apple: APPLE, sarah:SARAH}

I_s = {JACK, APPLE}

referential words

non-referential words

component in focus



scene:

Utterance: “Jack is eating the apple”

$P_R(w) = \gamma$

$P_{NR}(w) = 1-\gamma$

Model Design and Generative Process

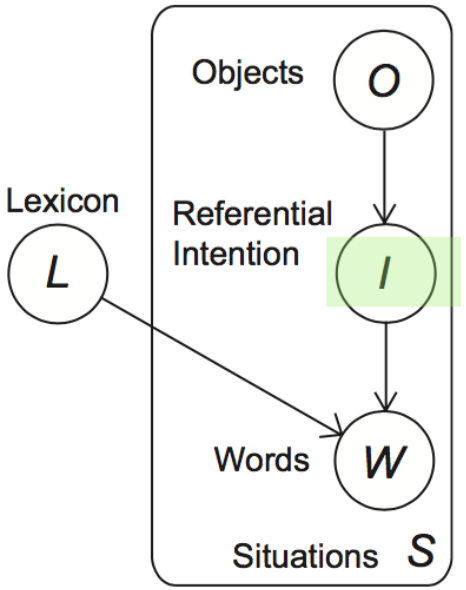
L = {Jack: JACK,apple: APPLE, sarah:SARAH}

I_S = {JACK, APPLE}

referential words

non-referential words

component in focus



scene:

Utterance: “Jack is eating the apple”

$P_R(w) = \gamma$

$P_{NR}(w) = 1-\gamma$

Model Design and Generative Process

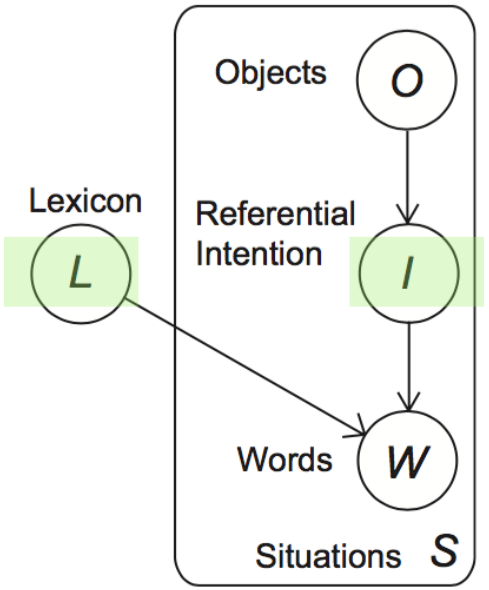
L = {Jack: JACK, apple: APPLE, sarah:SARAH}

I_S = {JACK, APPLE}

referential words

non-referential words

component in focus



scene:

Utterance: “Jack is eating the apple”

$P_R(w) = \gamma$

$P_{NR}(w) = 1-\gamma$

Model Design and Generative Process

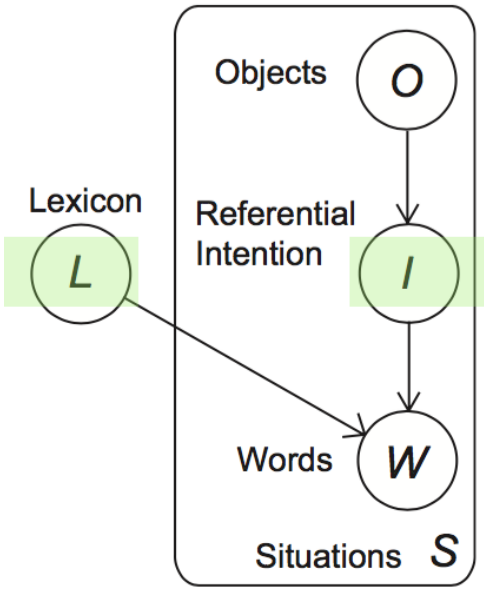
L = {Jack: JACK, apple: APPLE, sarah:SARAH}

I_S = {JACK, APPLE}

referential words

non-referential words

component in focus



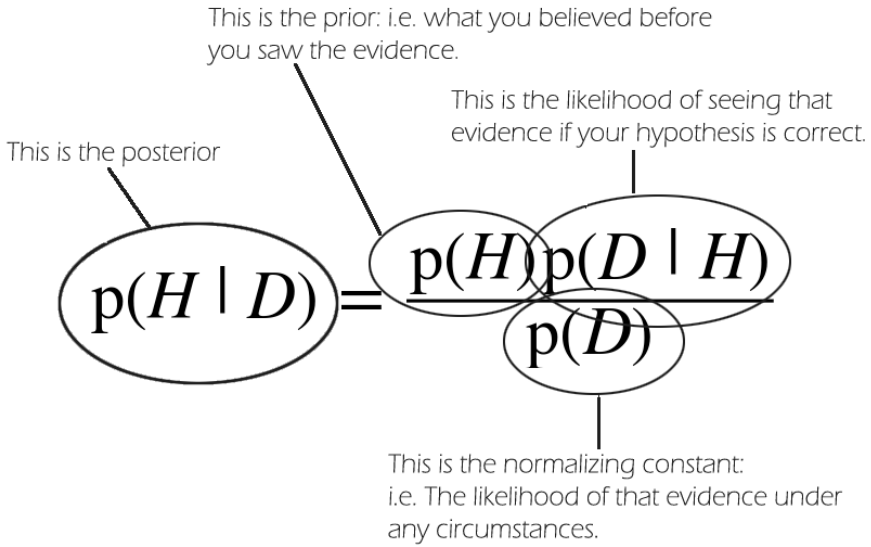
scene:

Utterance: "Jack is eating the apple"

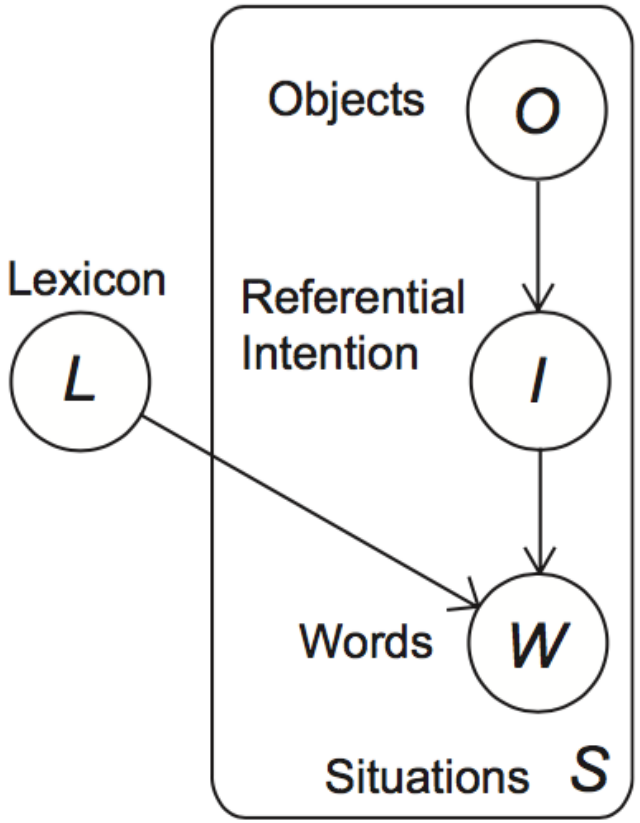
$P_R(w) = \gamma$

$P_{NR}(w) = 1-\gamma$

Reversing the Generative Process: Bayesian Inference



Posterior ∝ Likelihood × Prior



Bayesian Inference

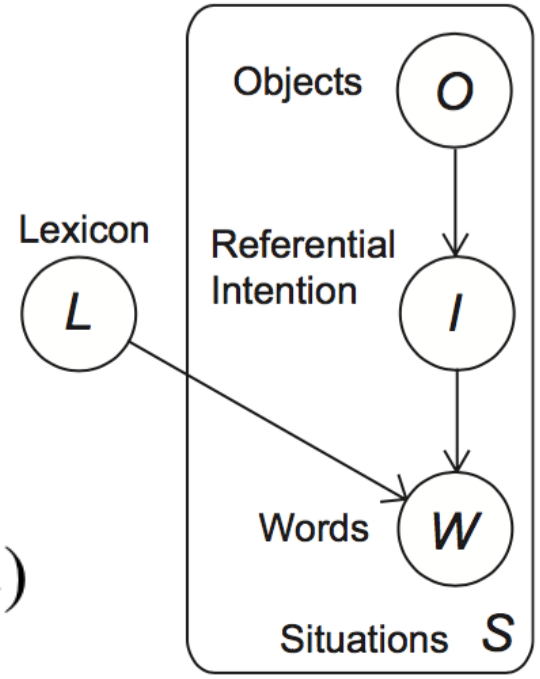
$$P(I_s|O_s) \propto 1$$

$$P(L) \propto e^{-\alpha \cdot |L|}$$

$$P(L|C) \propto P(C|L)P(L) \qquad (1)$$

$$P(C|L) = \prod_{s \in C} \sum_{I_s \subseteq O_s} P(W_s|I_s, L)P(I_s|O_s) \qquad (2)$$

$$P(W_s|I_s, L) = \prod_{w \in W_s} [\gamma \cdot \sum_{o \in I_s} \frac{1}{|I_s|} P_R(w|o, L) + (1-\gamma) P_{NR}(w|L)] \qquad (3)$$



Incremental and Memory-Limited Learning Algorithm

Model’s memory:

The knowledge in its lexicon and current situation.

Incremental and Memory-Limited Learning Algorithm

Incremental Word Learning:

(1) It only sees one situation at a time (no iteration over data).
(2) The model can only use the knowledge in its memory for hypothesis generation and hypothesis evaluation.
(3) The model maintains a single global lexicon (hypothesis) across all situations.
(4) The model makes local revisions to the global hypothesis by integrating an inferred mini-lexicon in the global hypothesis.
(5) Bayesian inference is only applied locally in the context of single situations based on context-appropriate word-referent pairs available in the memory (current lexicon and current situation)

Incremental Learning: Updating Lexicon

Inferring the MAP mini-lexicon in each situation:

(1) Generating mini-lexicon proposals (hypothesis generation)
using stochastic search techniques

(2) Scoring (hypothesis evaluation)
using relative posterior probability

Merging the new mini-lexicon with the current lexicon:

(1) Applying mutual exclusivity constraints to produce a preference for one-to-one mappings in the output lexicon.

Training and Evaluation Data

The evaluation data consists of 99 situations, with 33 unique situations repeated three times.

The repetition of the word learning situations is intentional to examine and determine the stability of the model as it received more inputs.

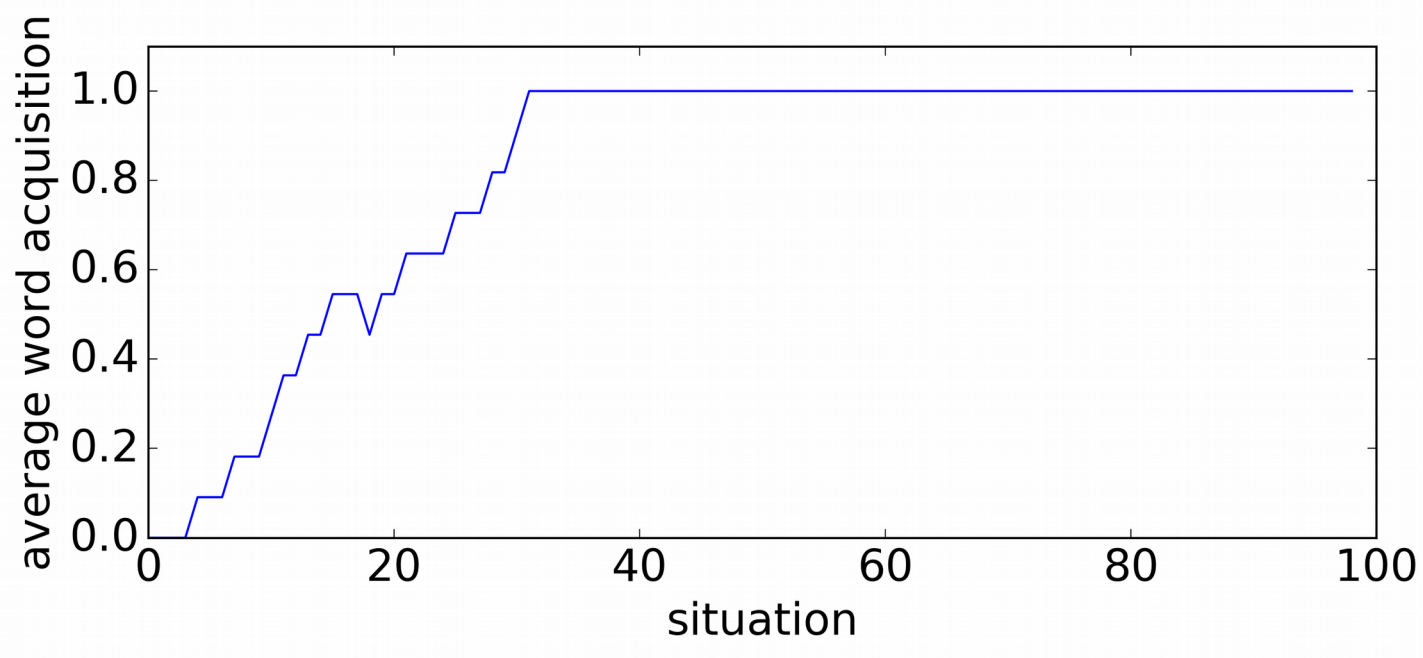
Utterance	Scene
bowl next to cup	BOWL,CUP,KNIFE
bowl next to cup	BOWL,CUP
look bowl	BOWL,KNIFE

Training and Evaluation Data

The evaluation data consists of 99 situations, with 33 unique situations repeated three times.

The repetition of the word learning situations is intentional to examine and determine the stability of the model as it received more inputs.

Utterance	Scene
bowl next to cup	BOWL,CUP,KNIFE
bowl next to cup	BOWL,CUP
look bowl	BOWL,KNIFE



Effects of Sensory Noise

We used our three parts of our DIARC architecture for evaluating the model's robustness to noise:

- a simulated speech recognition component
- a simulated visual object detection component
- a word learning component with the model implementation

To inject noise, we use *n*% chance that a "word" or "object" will be misclassified for each occurrence:

(1) **speech recognition noise:**
{0, 5, 10, 15, 20} % chance of misclassification

(2) **object recognition noise:**
{0, 5, 10, 15, 20} % chance of misclassification

Effects of Sensory Noise

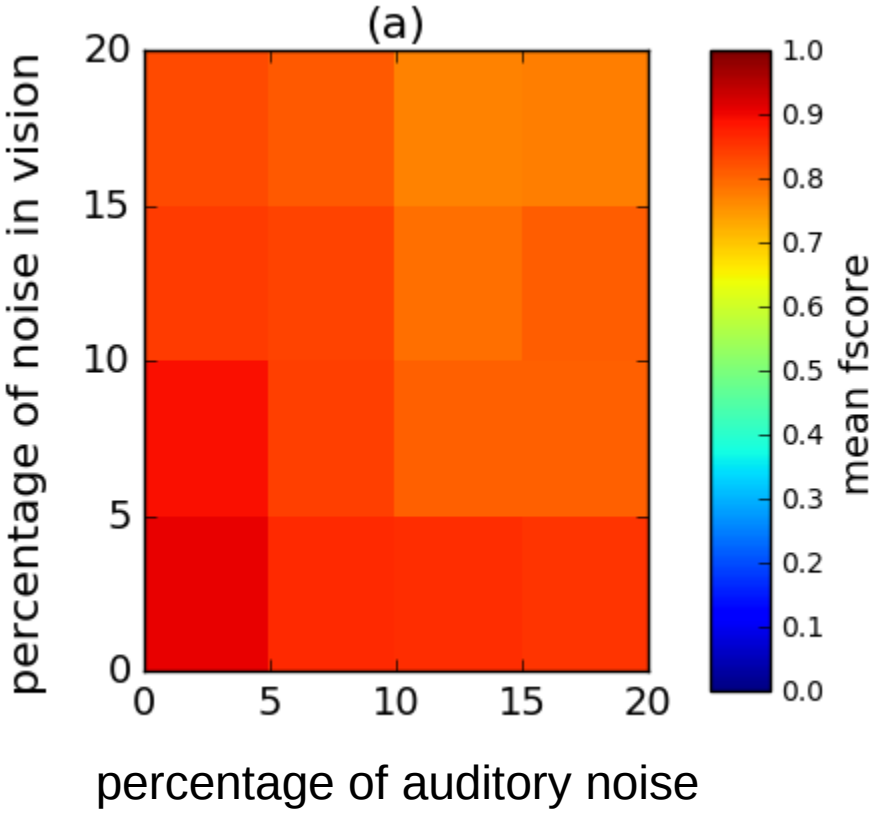
We used our three parts of our DIARC architecture for evaluating the model's robustness to noise:

- a simulated speech recognition component
- a simulated visual object detection component
- a word learning component with the model implementation

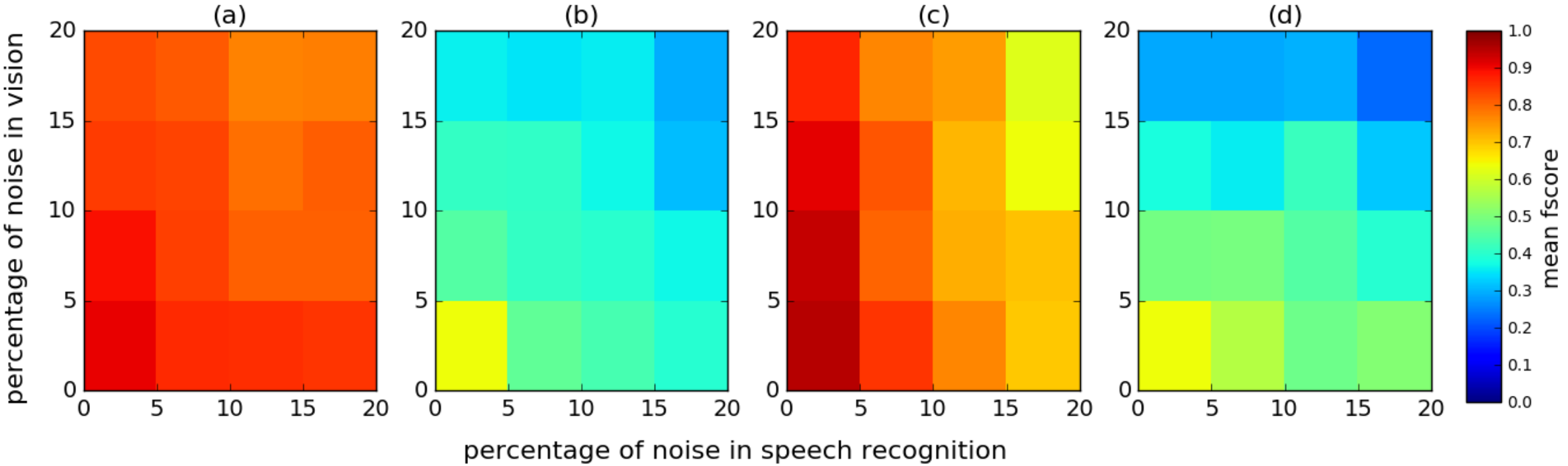
To inject noise, we use *n*% chance that a "word" or "object" will be misclassified for each occurrence:

(1) **speech recognition noise:**
{0, 5, 10, 15, 20} % chance of misclassification

(2) **object recognition noise:**
{0, 5, 10, 15, 20} % chance of misclassification

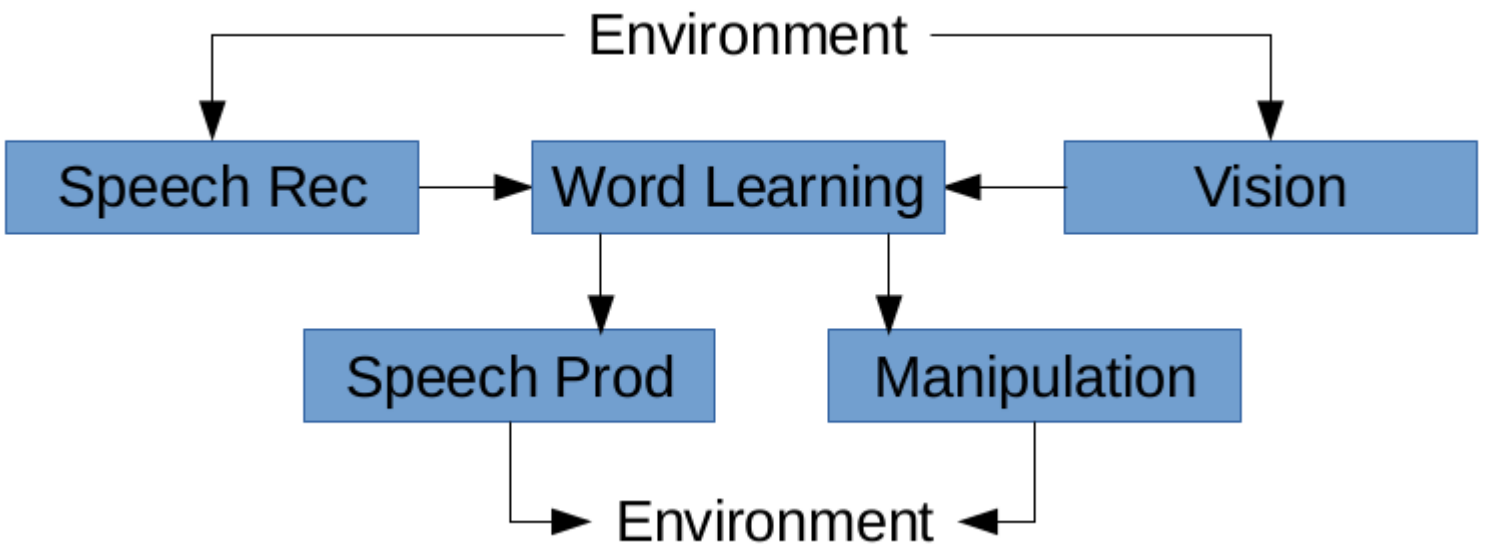


Comparison with Baseline Models



The heatmap of mean F-score values (averaged over 10 runs) for the lexica found by (a) our proposed incremental model, (b) the **association frequency model**, (c) the conditional probability **P(object|word)** model, and (d) the conditional probability **P(word|object)** model, under different noise conditions (as used by Frank et al. 2009)

Robot Proof-of-Concept Experiment



Components from thel DIARC architecture used in the proof-of-concept demonstration

Robot Proof-of-Concept Experiment

The interactions between human and robot:

(1) **during training:** robot reactions={<verbal: “OK”, motor: none>}

(2) **during testing:** robots reactions={<verbal: “here it is”, motor:point>,
<verbal: ”I don’t know what that is”, motor: none>}

https://hrilab.tufts.edu/movies/wordlearningdialogue.mp4 http://tiny.cc/68x5jy

Conclusion and Future Work

We presented an incremental and adaptive word learning model integrated into our cognitive robotic DIARC architecture and demonstrated how the model on a robot can learn new words in real-word settings.

The memory of our model is limited to the word-object mappings stored in the lexicon and the single situation is sees at each point in time.

The model exhibits superior performance and robustness to noise in comparison with the baseline incremental models.

In related work, we have recently extended the model to include the joint acquisition of simple word order (syntax) in conjunction with semantics (Sadeghi and Scheutz, 2017)

In a next step, we are planning to extend this model to verb acquisition.