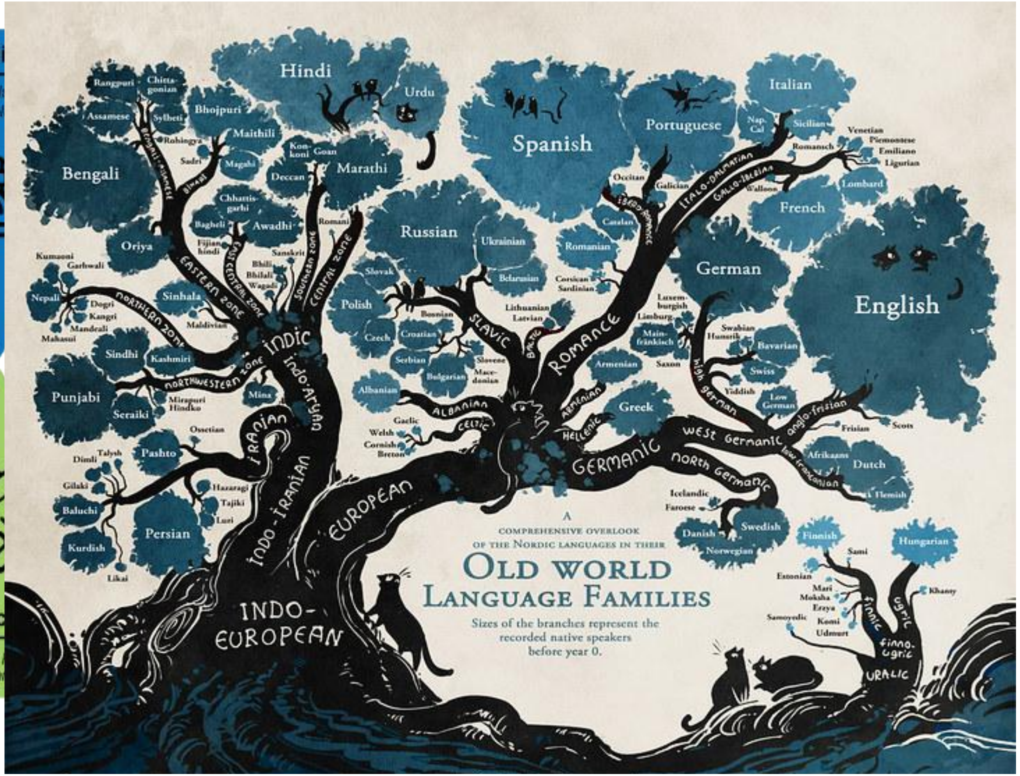
Categories and Instances in Human Cognition and AI

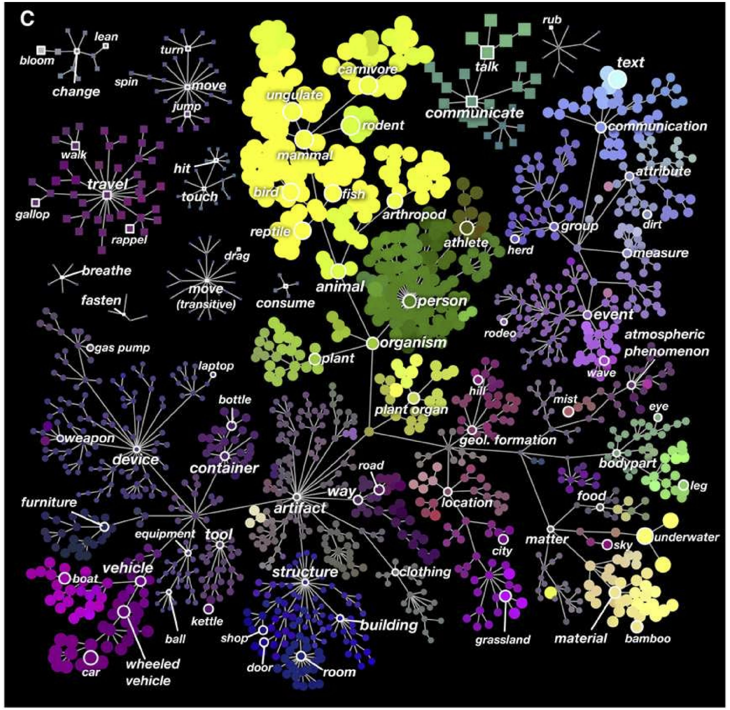
Aida Nematzadeh
DeepMind

Categories Are Everywhere



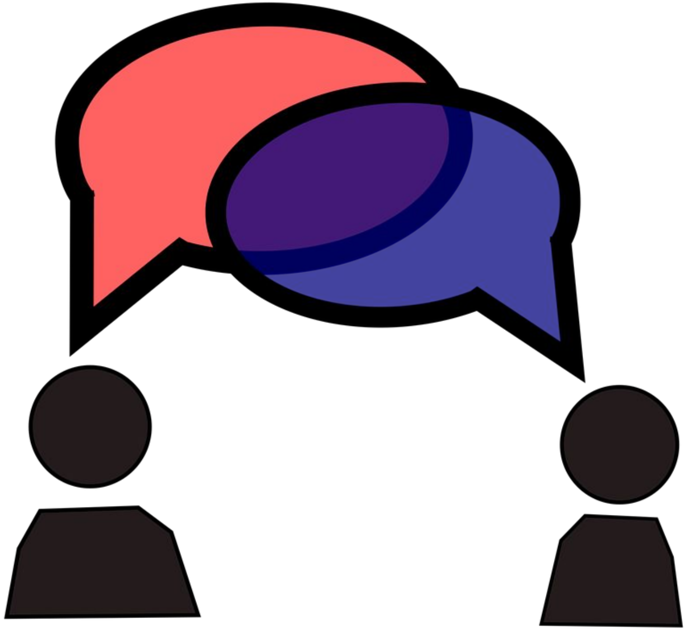
Categories Enable Generalization

Categories Facilitate Search



[Huth *et al.*, Neuron 2012]

But We Also Rely on Specific Instances



AI Systems Need a Similar Capacity

- Form categories at different levels of abstractions
- Represent and reason about instances

AI Systems Need a Similar Capacity

- Form categories at different levels of abstractions → novel word generalization

- Represent and reason about instances → theory of mind

AI Systems Need a Similar Capacity

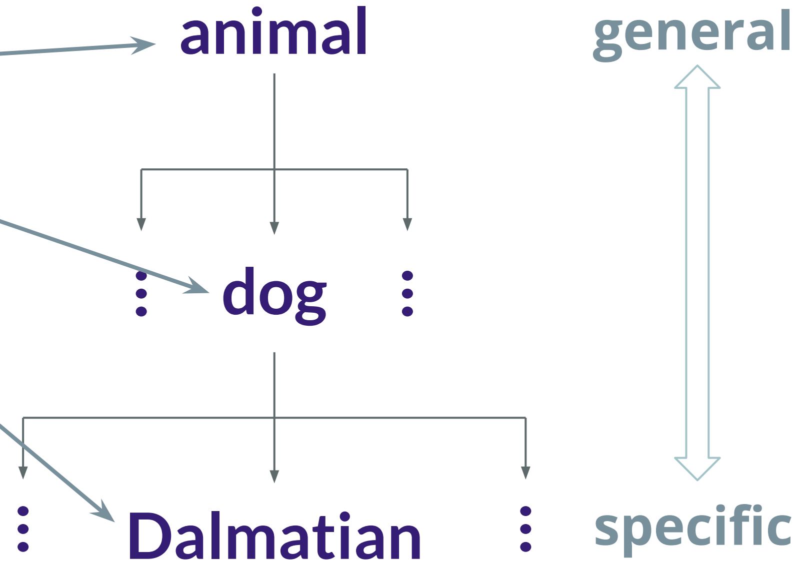
- Form categories at different levels of abstractions → novel word generalization

- Represent and reason about instances → theory of mind

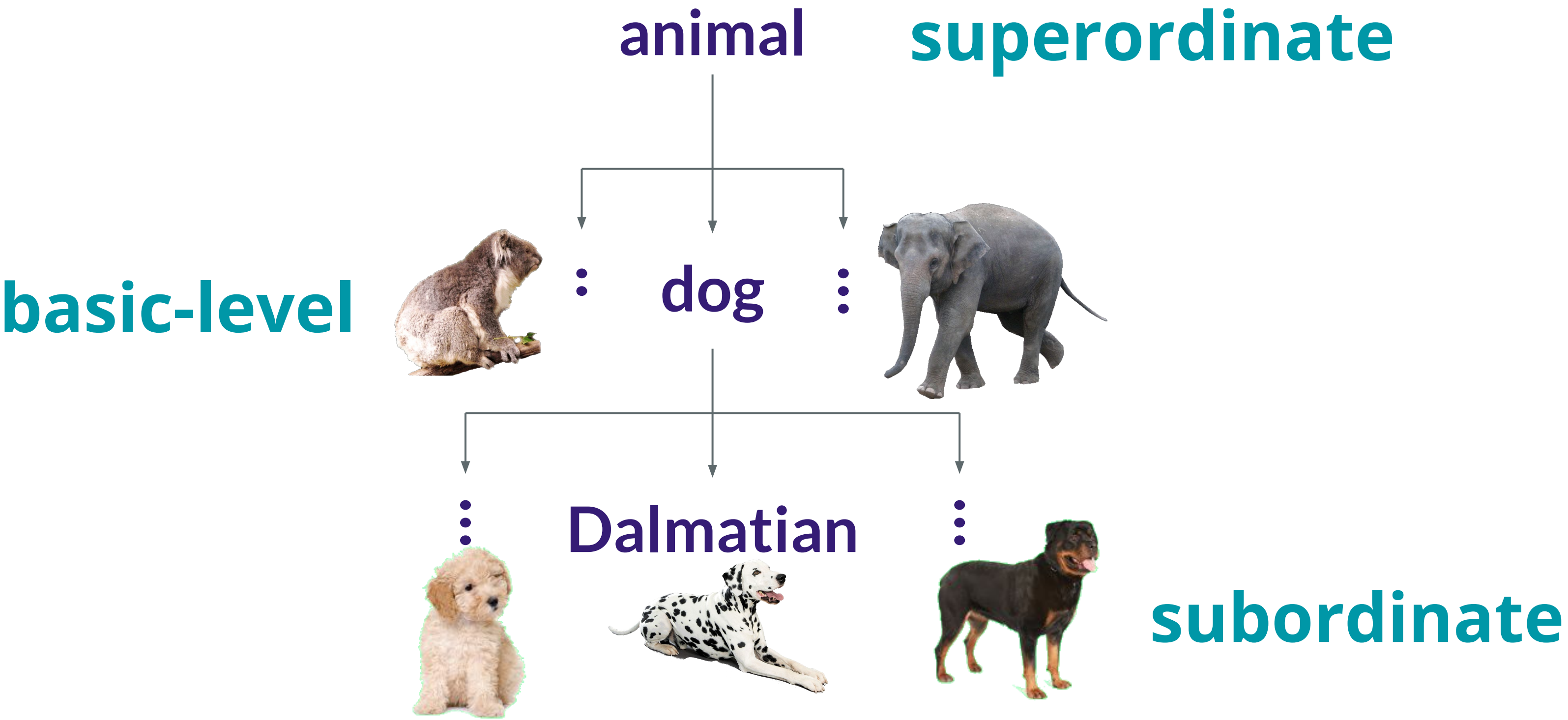
Novel Word Generalization

To which level of a hierarchical taxonomy does a word refer?

"dax"



Different Levels of a Taxonomy



Generalization in People

[Xu & Tenenbaum, Psych Rev 2007]

How to generalize words from a few examples?

Train (3 sub)

trial 1 *This is a dax.*

trial 2 *Here is a dax.*

trial 3 *A dax.*

Test

Pick everything that is a dax

Generalization in People

[Xu & Tenenbaum, Psych Rev 2007]

How to generalize words from a few examples?

Train (3 basic)

trial 1 *This is a dax.*

trial 2 *Here is a dax.*

trial 3 *A dax.*

Test

Pick everything that is a dax

Generalization in People

[Xu & Tenenbaum, Psych Rev 2007]

How to generalize words from a few examples?

Train (1 sub)

trial 1 *This is a dax.*

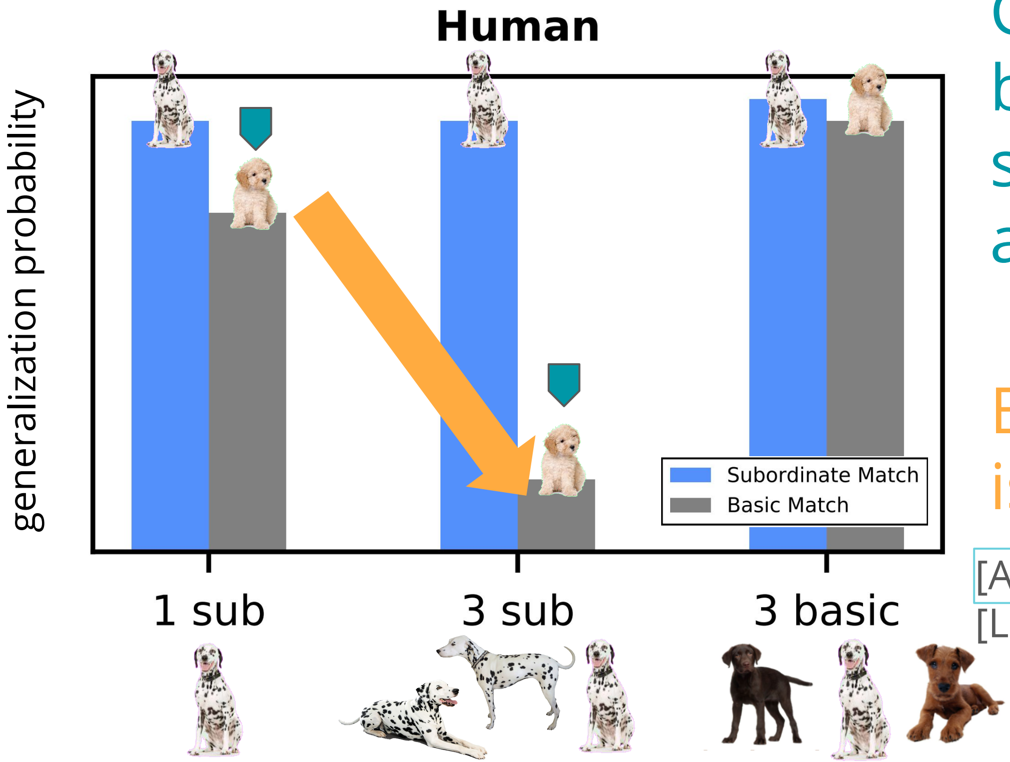
Test

Pick everything that is a dax

Generalization in People

[Xu & Tenenbaum, Psych Rev 2007]

basic-level generalization



Generalize to the basic-level with only subordinate examples: a basic-level bias.

Basic-level generalization is attenuated.

[Abbott, Austerweil, & Griffiths, CogSci 2012]
[Lewis & Frank, Psych Sci 2018]

Why Are the Results Interesting?

People learn a novel word (“dax”) only from positive examples.

They exhibit a bias towards the basic-level category: is this bias learned or innate?

Their generalization is sensitive to the number of examples in a category.

What Does it Take for a Model to Generalize Novel Words?

Xu & Tenenbaum, Psych Rev 2007 [paper]

- A Bayesian Model:

h is a hypothesis about the novel word's meaning; e.g., all dogs

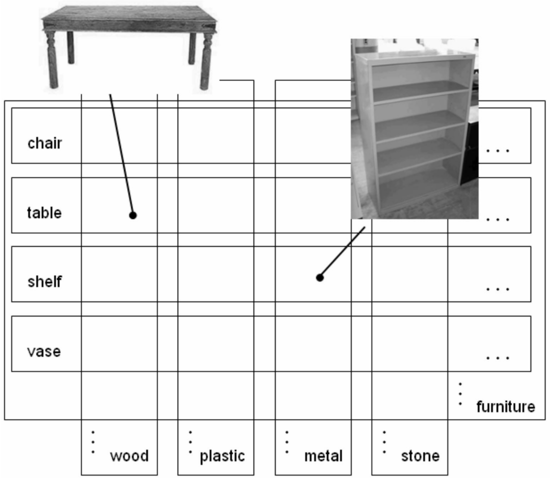
$$p(h \mid X) = \frac{p(X \mid h)\, p(h)}{p(X)}$$

X is the set of observations; e.g., 3 Dalmatians

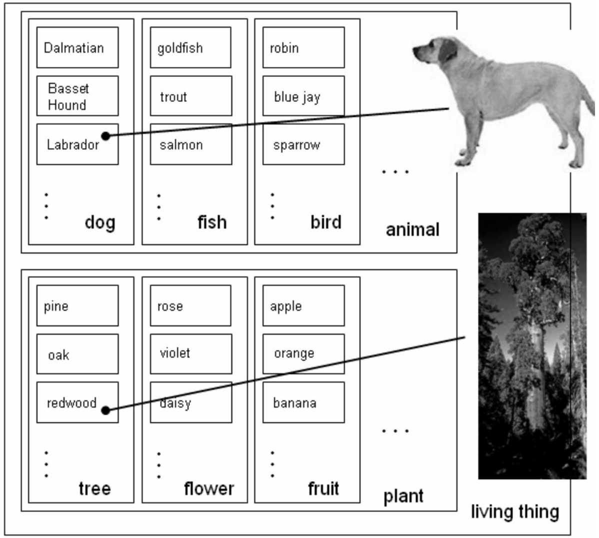
- Need to define a **hypothesis space *h***, the **likelihood *P(X|h)***, and the **prior *p(h)***.

Xu & Tenenbaum, Psych Rev 2007 [paper]

Possible hypothesis spaces:



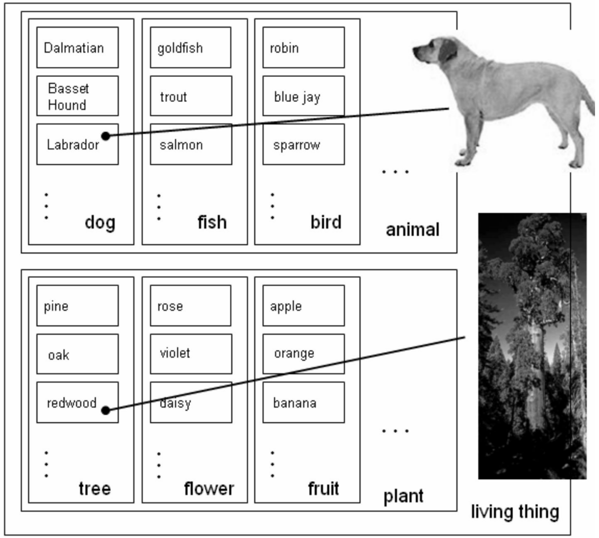
Feature overlap



Nested categories

Xu & Tenenbaum, Psych Rev 2007[paper]

The prior *p(h)* assigns zero to any hypothesis not valid given this taxonomy (nested categories):



Xu & Tenenbaum, Psych Rev 2007 [paper]

The likelihood:

$$p(X \mid h) = \left[\frac{1}{\text{size}(h)}\right]^n$$

Smaller categories are prefered & exponentially so as the number of observations increase → **size principle**

Encodes a lot of knowledge about the taxonomy and the generalization mechanism.

A K-Shot Generalization Task for AI Models

Xu & Tenenbaum, Psych Rev 2007:

- Model is trained and tested on the task.
- Size principle: smaller categories are prefered to the larger ones and exponentially so as the number of observations increase.

Nematzadeh et al., EMNLP 2015[paper]

An alignment-based word learning model:

1) Align features to a word given what the model has learned.

Utterance: Look at the Dalmatian.

Scene: { LOOK, DALMATIAN, DOG, ANIMAL }

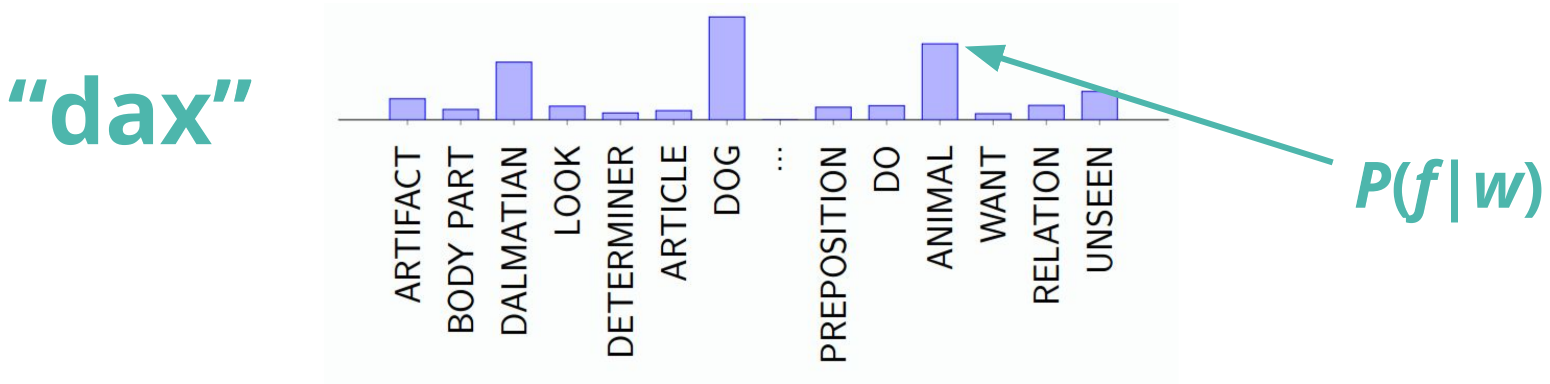
2) Update the model's knowledge based on these alignments.

Utterance: Look at the Dalmatian.

Scene: { LOOK, DALMATIAN, DOG, ANIMAL }

Nematzadeh et al., EMNLP 2015[paper]

An alignment-based word learning model: Given a set of utterance-scene pairs, learns a meaning representations for each word, P(.|w):



Nematzadeh et al., EMNLP 2015 [paper]

1) Align features to a word given what the model has learned, $p(.|w)$.

Utterance: Look at the Dalmatian.

Scene: { LOOK, DALMATIAN, DOG, ANIMAL }

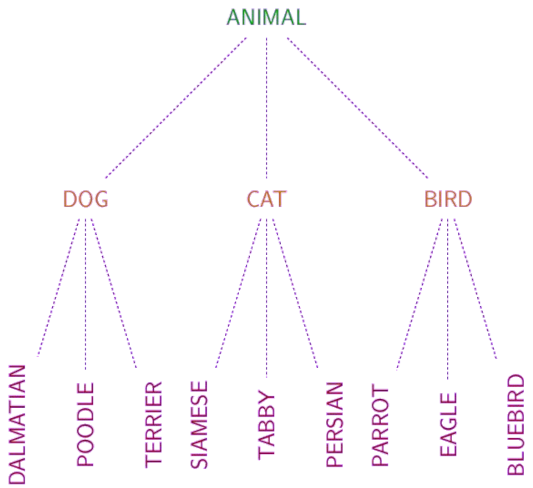
2) Update the model's knowledge based on these alignments.

Utterance: Look at the Dalmatian.

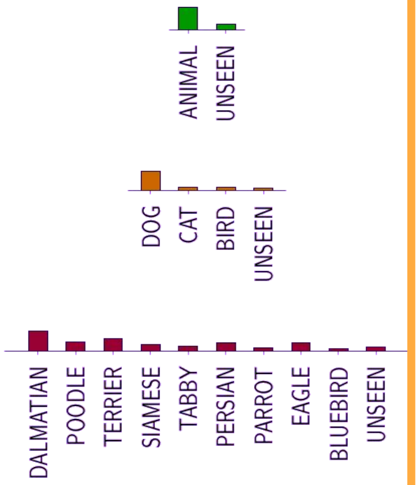
Scene: { LOOK, DALMATIAN, DOG, ANIMAL }

Nematzadeh et al., EMNLP 2015[paper]

people's taxonomic knowledge



the model learns



meaning of "dax"

the model assumes

group $\mathcal{G}_1$

group $\mathcal{G}_2$

group $\mathcal{G}_3$

Nematzadeh et al., EMNLP 2015[paper]

Generalization should be influenced by both **token** and **type** frequency.

3 types
3 tokens

1 type
3 tokens

$$P_t(f|w) = \frac{\mathrm{assoc}_t(f, w) + \lambda_{\mathcal{G}}}{\sum_{f' \in \mathcal{G}} \mathrm{assoc}_t\underline{(f', w)} + \beta_{\mathcal{G}} \times \lambda_{\mathcal{G}}}$$

Prior reflects function of number of types in a group

Likelihood reflects the token frequency of observed features

A K-Shot Generalization Task for AI Models

Nematzadeh *et al.*, EMNLP 2015:

- An alignment-based translation model; *tested* on the novel word generalization task.
- Both token and type frequencies influence generalization.

3 types
3 tokens

1 type
3 tokens

Peterson et al., CogSci 2018[paper]

A multi-label image classification model; tested on the novel word generalization task.

Introduced the size principle in the inference procedure.

A K-Shot Generalization Task for AI Models

Peterson *et al.*, CogSci 2018:

- A multi-label image classification model; tested on the novel word generalization task.
- Introduced the size principle in the inference procedure.

Grant et al., CogSci 2019[paper]

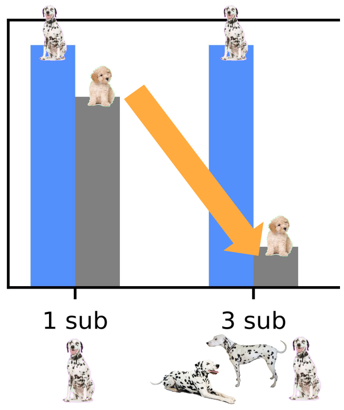
Formulate the task as predicting a binary label for an input; the label determines if the input belongs to a concept (*e.g.*, all dogs).

Propose a meta learning approach to estimate decision-boundaries for each concept.

Grant et al., CogSci 2019[paper]

- Uses a sampling approach that assumes knowledge of hierarchical taxonomy: negative examples are drawn from other concepts.

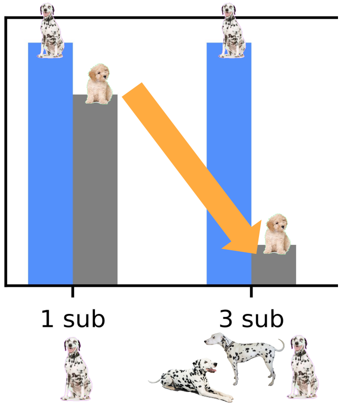
- Does not replicate the decrease in the basic-level generalization.



A K-Shot Generalization Task for AI Models

Grant et al., CogSci 2019:

- A meta learning approach to estimate decision-boundaries from only positive examples.
- Does not replicate the decrease in the basic-level generalization.



A K-Shot Generalization Task for AI Models

	input data	encoded knowledge
Xu & Tenenbaum, 2007	Artificial data.	Hierarchical taxonomy. Size principle
Nematzadeh *et al.*, 2015	Natural sentences. Symbols to represent scenes ("images")	Feature groups. Number of types in a feature group.
Peterson *et al.*, 2018	Word labels. Natural images.	Hierarchical taxonomy. Size principle
Grant et al., 2019	Word labels. Natural images.	Hierarchical taxonomy.

Can we reduce the amount of encoded knowledge?

AI Systems Need a Similar Capacity

- Form categories at different levels of abstractions → novel word generalization; current models need biases sensitive to the number/size of instances/categories.
- Represent and reason about instances → theory of mind.

AI Systems Need a Similar Capacity

- Form categories at different levels of abstractions → novel word generalization; current models need biases sensitive to the number/size of instances/categories.
- Represent and reason about instances → theory of mind.

Remembering and Representing Instances

Mary got the milk there.
Sandra went back to the kitchen.
Mary travelled to the hallway.

Q: Where is Mary? A: hallway

Q: Where is the milk? A: hallway

The bAbi Dataset of Reasoning

[Weston *et al.*, ICLR 2016]

20 different types of reasoning tasks:

The last sentence has the answer.

Task 1: Single Supporting Fact
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Current models fail only a few of the bAbi tasks.

Do models answer a question using the *right* information?

The bAbi Dataset of Reasoning

[Weston *et al.*, ICLR 2016]

20 different types of reasoning tasks:

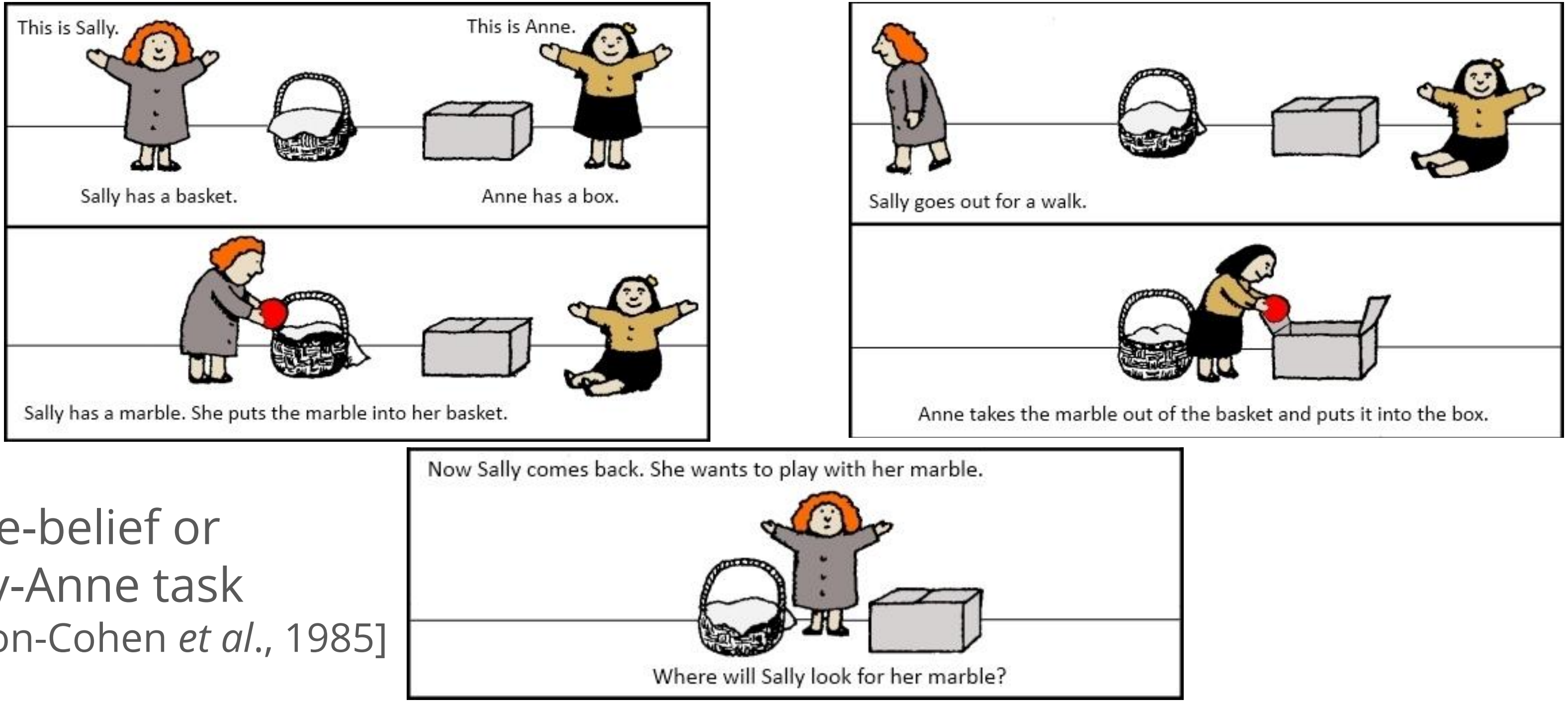
Task 1: Single Supporting Fact
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Current models fail only a few of the bAbi tasks.

Do models answer a question using the *right* information?

Theory of Mind: Reasoning About Beliefs



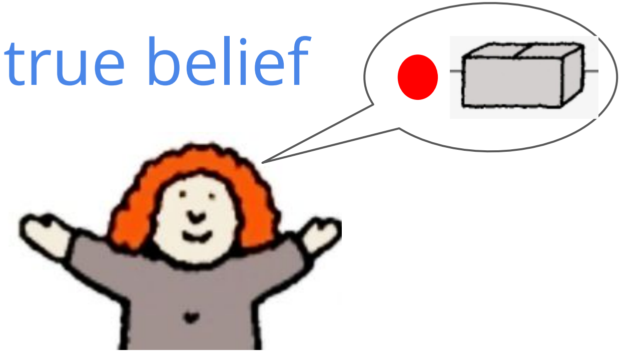
False-belief or
Sally-Anne task
[Baron-Cohen *et al.*, 1985]

Need to reason about others' beliefs & maintain multiple representations.

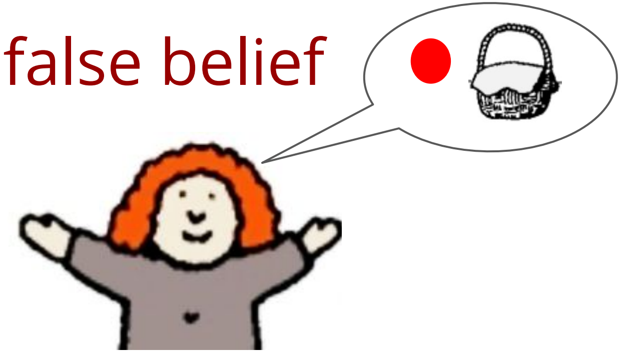
True or False Beliefs

reality

true belief

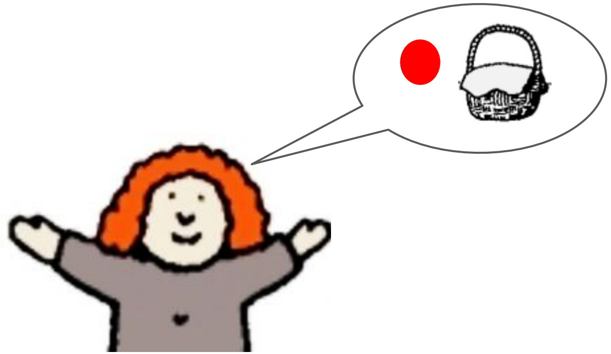


false belief



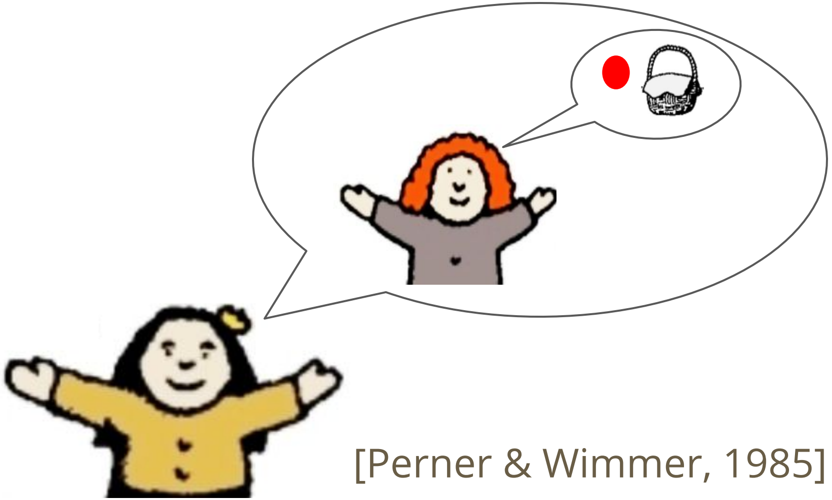
Beliefs About Beliefs

First-order belief: Sally's belief about marble's location.



false

Second-order belief: Anne's belief about Sally's belief.



[Perner & Wimmer, 1985]

True Belief

Anne entered the kitchen.
Sally entered the kitchen.
The milk is in the fridge.
Anne moved the milk to the pantry.

Memory Where was the milk at the beginning?
Reality Where is the milk really?
First-order Where will Sally look for the milk?
Second-order Where does Anne think that Sally searches for the milk?

[Nematzadeh *et al.*, EMNLP 2018]

False Belief

Anne entered the kitchen.
Sally entered the kitchen.
The milk is in the fridge.
Sally exited the kitchen.
Anne moved the milk to the pantry.

Memory	Where was the milk at the beginning?
Reality	Where is the milk really?
First-order	Where will Sally look for the milk?
Second-order	Where does Anne think that Sally searches for the milk?

[Nematzadeh *et al.*, EMNLP 2018]

Second-order False Belief

Anne entered the kitchen.
Sally entered the kitchen.
The milk is in the fridge.
Sally exited the kitchen.
Anne moved the milk to the pantry.
Anne exited the kitchen.
Sally entered the kitchen.

Memory	Where was the milk at the beginning?
Reality	Where is the milk really?
First-order	Where will Sally look for the milk?
Second-order	Where does Anne think that Sally searches for the milk?

[Nematzadeh *et al.*, EMNLP 2018]

Theory of Mind Tasks

		tasks	
	True Belief	**False Belief**	**Second-order False Belief**
questions **Memory**	fridge	fridge	fridge
Reality	pantry	pantry	pantry
First-order	pantry	fridge	pantry
Second-order	pantry	fridge	fridge

We group 5 task-question pairs to form a story.

Evaluating Memory-Augmented Models

End-to-End Memory Nets [Sukhbaatar *et al.*, 2015]

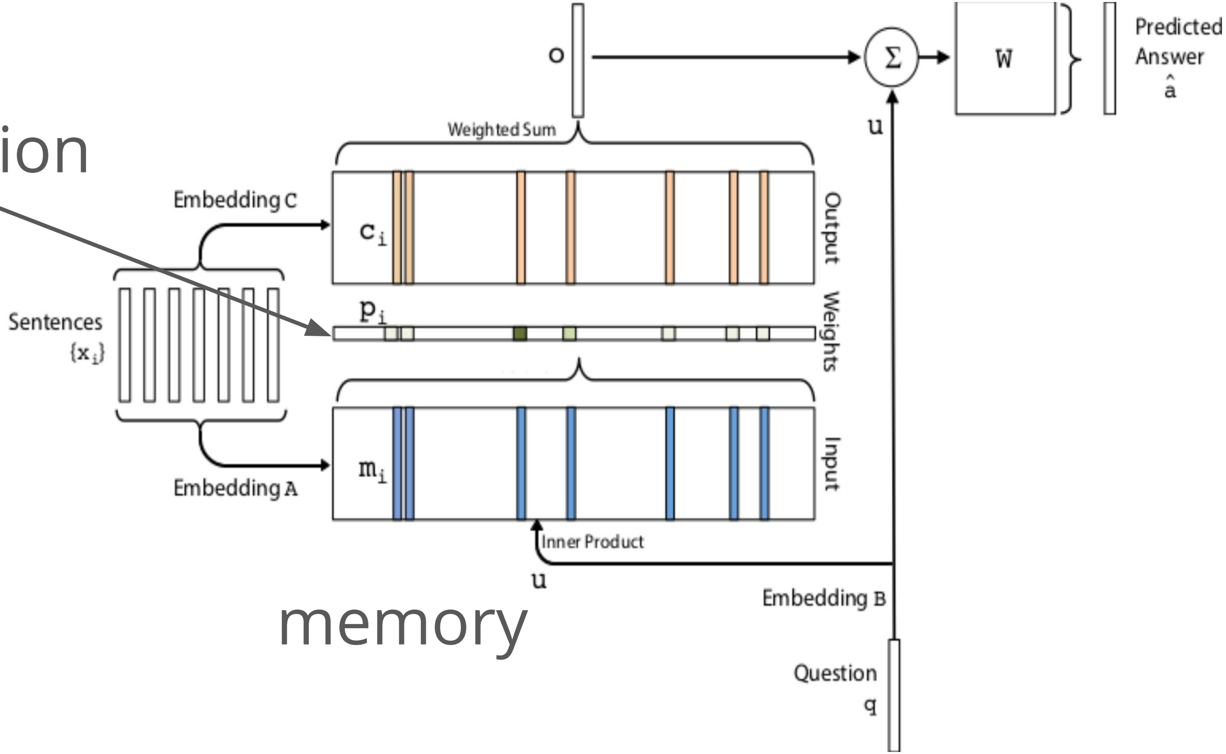
Multiple Observer Model [Grant *et al.*, 2017]

Recurrent Entity Networks [Henaff *et al.*, 2017]

Relation Networks [Santoro *et al.*, 2017]

End-to-End Memory Nets [Sukhbaatar et *al*., 2015]

attention



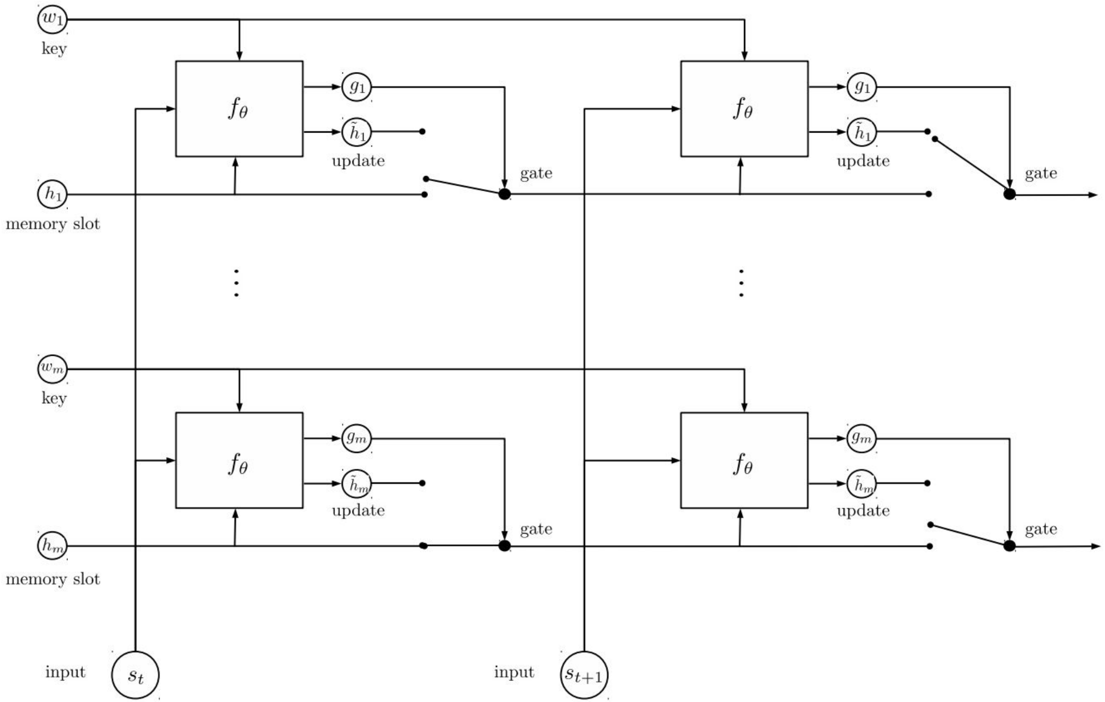
memory

Multiple Observer Model [Grant et al., 2017]

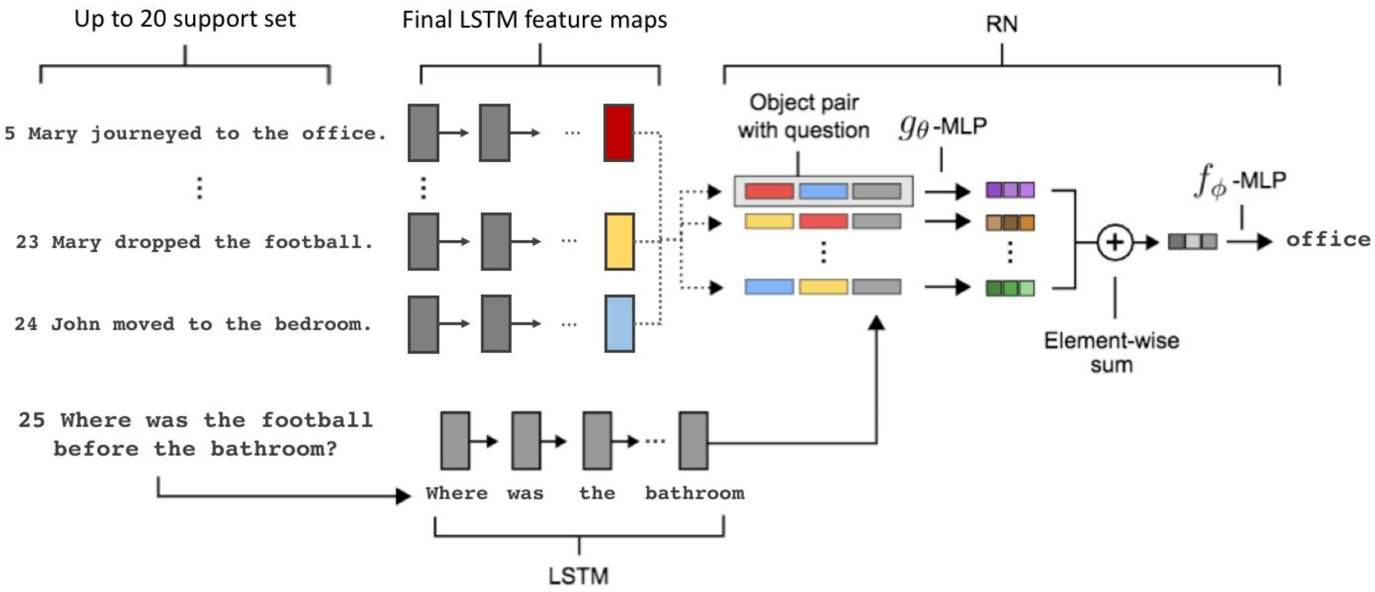
Extends **MemN2N** to to have separate memories for Sally, Anne, and the observer.

Adds attention over these memories.

Recurrent Entity Networks [Henaff et al., 2017]



Relation Networks [Santoro et al., 2017]



Results: Hardest Questions

models / tasks	True Belief	False Belief	Second-order False Belief
MemN2N [Sukhbaatar et al., 2015]	Second-order Belief (42.9)	First-order Belief (17.3)	First-order Belief (56.4)
Multiple Observer [Grant et al., 2017]	Memory (93.2)	First-order Belief (56.4)	First- & Second-order Belief (90.3)

First-order belief questions are harder than the second-order ones.

Why First-order Beliefs Are Harder?

	True Belief	False Belief	Second-order False Belief
First-order	pantry	**fridge**	**pantry**
Second-order	pantry	fridge	fridge

The answer to the first-order question is **not** the same for the two similar tasks.

Results: Hardest Questions

models / tasks	True Belief	False Belief	Second-order False Belief
MemN2N [Sukhbaatar et al., 2015]	Second-order Belief (42.9)	First-order Belief (17.3)	First-order Belief (56.4)
Multiple Observer [Grant et al., 2017]	Memory (93.2)	First-order Belief (56.4)	First- & Second-order Belief (90.3)
EntNet [Henaff et al., 2017]	Memory (74.0)	Memory (76.1)	Memory (74.3)

Results: Hardest Questions

models / tasks	True Belief	False Belief	Second-order False Belief
MemN2N [Sukhbaatar et al., 2015]	Second-order Belief (42.9)	First-order Belief (17.3)	First-order Belief (56.4)
Multiple Observer [Grant et al., 2017]	Memory (93.2)	First-order Belief (56.4)	First- & Second-order Belief (90.3)
EntNet [Henaff et al., 2017]	Memory (74.0)	Memory (76.1)	Memory (74.3)
RelNet [Santoro et al., 2017]	Memory (79.2)	Memory (77.9)	Memory (77.7)

Summary of Results

Models **with** explicit memory (MemN2N and Multiple Observer) fail at belief questions.

But EntNet and RelNet fail at the memory questions.

Results: Experimenting with Noise

Introduce “noise” sentences randomly.

Anne entered the kitchen
Sally entered the kitchen. → Phone rang.
The milk is in the fridge.
Sally exited the kitchen.
Anne moved the milk to the pantry.

Performance of all models decrease -- they are not using the right information.

Representing Categories and Instances

AI models need to represent categories at different levels of abstraction.

They also need to represent and *remember* important *instances*.

Experiments in developmental psychology provide interesting framework for evaluating AI models.

Thanks!

Acknowledgments

Kaylee Burns
Stanford

Erin Grant
UC Berkeley

Tom Griffiths
Princeton

Alison Gopnik
UC Berkeley

Josh Peterson
Princeton

Paul Soulos
JHU

Suzanne Stevenson
U of Toronto