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"

animal

:: dog ::

:: Dalmatian ::

general

specific
Different Levels of a Taxonomy

animal

superordinate

basic-level

dog

Dalmatian

subordinate
Generalization in People

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

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

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

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?
A Bayesian Model:

- $h$ is a hypothesis about the novel word’s meaning; e.g., all dogs
- $X$ is the set of observations; e.g., 3 Dalmatians

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

- Need to define a hypothesis space $h$, the likelihood $P(X \mid h)$, and the prior $p(h)$. 
Possible hypothesis spaces:
The prior $p(h)$ assigns zero to any hypothesis not valid given this taxonomy (nested categories):
The likelihood:

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

Smaller categories are preferred & exponentially so as the number of observations increase $\rightarrow$ **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 preferred to the larger ones and exponentially so as the number of observations increase.
Nematzadeh et al., EMNLP 2015 \[\text{[paper]}\]

An alignment-based word learning model:

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

\[
\begin{align*}
\text{Utterance:} & \quad \text{Look at the Dalmatian.} \\
\text{Scene:} & \quad \{ \text{LOOK, DALMATIAN, DOG, ANIMAL} \}
\end{align*}
\]

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

\[
\begin{align*}
\text{Utterance:} & \quad \text{Look at the Dalmatian.} \\
\text{Scene:} & \quad \{ \text{LOOK, DALMATIAN, DOG, ANIMAL} \}
\end{align*}
\]
An alignment-based word learning model: Given a set of utterance-scene pairs, learns a meaning representations for each word, \( P(. \mid w) \):

“dax”

\[ P(f \mid w) \]
1) Align features to a word given what the model has learned, $p(. | w)$.

**Utterance:** Look at the Dalmatian.

**Scene:** { LOOK, **D**ALMATIAN, DOG, **A**NIMAL }

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

**Utterance:** Look at the Dalmatian.

**Scene:** { LOOK, **D**ALMATIAN, DOG, **A**NIMAL }
people’s taxonomic knowledge

the model learns

the model assumes

meaning of “dax”

Dalmatian

Poodle

Terrier

Siamese

Tabby

Persian

Parrot

Eagle

Bluebird

Animal

Unseen

Group $G_1$

Group $G_2$

Group $G_3$
Generalization should be influenced by both token and type frequency.

\[ P_t(f|w) = \frac{\text{assoc}_t(f, w) + \lambda g}{\sum_{f' \in G} \text{assoc}_t(f', w) + \beta g \times \lambda g} \]

Likelihood reflects the token frequency of observed features

Prior reflects function of number of types in a group

Nematzadeh et al., EMNLP 2015 [paper]
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.
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.
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

<table>
<thead>
<tr>
<th>input data</th>
<th>encoded knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xu &amp; Tenenbaum, 2007</td>
<td>Artificial data.</td>
</tr>
<tr>
<td>Nematzadeh et al., 2015</td>
<td>Natural sentences. Symbols to represent scenes (“images”)</td>
</tr>
<tr>
<td>Peterson et al., 2018</td>
<td>Word labels. Natural images.</td>
</tr>
</tbody>
</table>

**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

20 different types of reasoning tasks:

The last sentence has the answer.

Current models fail only a few of the bAbi tasks.

Do models answer a question using the right information?

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

[Weston et al., ICLR 2016]
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

true belief

false belief
Beliefs About Beliefs

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

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

[False]

[Perner & Wimmer, 1985]
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]
<table>
<thead>
<tr>
<th>False Belief</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anne entered the kitchen.</td>
</tr>
<tr>
<td>Sally entered the kitchen.</td>
</tr>
<tr>
<td>The milk is in the fridge.</td>
</tr>
<tr>
<td><strong>Sally exited the kitchen.</strong></td>
</tr>
<tr>
<td>Anne moved the milk to the pantry.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where was the milk at the beginning?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where is the milk really?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>First-order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where will Sally look for the milk?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Second-order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where does Anne think that Sally searches for the milk?</td>
</tr>
</tbody>
</table>

[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

<table>
<thead>
<tr>
<th>questions</th>
<th>True Belief</th>
<th>False Belief</th>
<th>Second-order False Belief</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>fridge</td>
<td>fridge</td>
<td>fridge</td>
</tr>
<tr>
<td>Reality</td>
<td>pantry</td>
<td>pantry</td>
<td>pantry</td>
</tr>
<tr>
<td>First-order</td>
<td>pantry</td>
<td>fridge</td>
<td>pantry</td>
</tr>
<tr>
<td>Second-order</td>
<td>pantry</td>
<td>fridge</td>
<td>fridge</td>
</tr>
</tbody>
</table>

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]
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

<table>
<thead>
<tr>
<th>models</th>
<th>tasks</th>
<th>True Belief</th>
<th>False Belief</th>
<th>Second-order False Belief</th>
</tr>
</thead>
<tbody>
<tr>
<td>MemN2N [Sukhbaatar et al., 2015]</td>
<td>Second-order Belief (42.9)</td>
<td>First-order Belief (17.3)</td>
<td>First-order Belief (56.4)</td>
<td></td>
</tr>
<tr>
<td>Multiple Observer [Grant et al., 2017]</td>
<td>Memory (93.2)</td>
<td>First-order Belief (56.4)</td>
<td>First- &amp; Second-order Belief (90.3)</td>
<td></td>
</tr>
</tbody>
</table>

First-order belief questions are harder than the second-order ones.
Why First-order Beliefs Are Harder?

<table>
<thead>
<tr>
<th></th>
<th>True Belief</th>
<th>False Belief</th>
<th>Second-order False Belief</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First-order</strong></td>
<td>pantry</td>
<td>fridge</td>
<td>pantry</td>
</tr>
<tr>
<td><strong>Second-order</strong></td>
<td>pantry</td>
<td>fridge</td>
<td>fridge</td>
</tr>
</tbody>
</table>

The answer to the first-order question is **not** the same for the two similar tasks.
## Results: Hardest Questions

<table>
<thead>
<tr>
<th>models</th>
<th>True Belief</th>
<th>False Belief</th>
<th>Second-order False Belief</th>
</tr>
</thead>
<tbody>
<tr>
<td>MemN2N [Sukhbaatar et al., 2015]</td>
<td>Second-order Belief (42.9)</td>
<td>First-order Belief (17.3)</td>
<td>First-order Belief (56.4)</td>
</tr>
<tr>
<td>Multiple Observer [Grant et al., 2017]</td>
<td>Memory (93.2)</td>
<td>First-order Belief (56.4)</td>
<td>First- &amp; Second-order Belief (90.3)</td>
</tr>
<tr>
<td>EntNet [Henaff et al., 2017]</td>
<td>Memory (74.0)</td>
<td>Memory (76.1)</td>
<td>Memory (74.3)</td>
</tr>
</tbody>
</table>
## Results: Hardest Questions

<table>
<thead>
<tr>
<th>models</th>
<th>tasks</th>
<th>True Belief</th>
<th>False Belief</th>
<th>Second-order False Belief</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MemN2N</strong> [Sukhbaatar et al., 2015]</td>
<td>Second-order Belief (42.9)</td>
<td>First-order Belief (17.3)</td>
<td>First-order Belief (56.4)</td>
<td></td>
</tr>
<tr>
<td><strong>Multiple Observer</strong> [Grant et al., 2017]</td>
<td>Memory (93.2)</td>
<td>First-order Belief (56.4)</td>
<td>First- &amp; Second-order Belief (90.3)</td>
<td></td>
</tr>
<tr>
<td><strong>EntNet</strong> [Henaff et al., 2017]</td>
<td>Memory (74.0)</td>
<td>Memory (76.1)</td>
<td>Memory (74.3)</td>
<td></td>
</tr>
<tr>
<td><strong>RelNet</strong> [Santoro et al., 2017]</td>
<td>Memory (79.2)</td>
<td>Memory (77.9)</td>
<td>Memory (77.7)</td>
<td></td>
</tr>
</tbody>
</table>
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.
The milk is in the fridge.
*Sally exited the kitchen.*
Anne moved the milk to the pantry.

Phone rang.

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.
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