Learning, Representing, and Understanding Language in Humans and Computers

Aida Nematzadeh

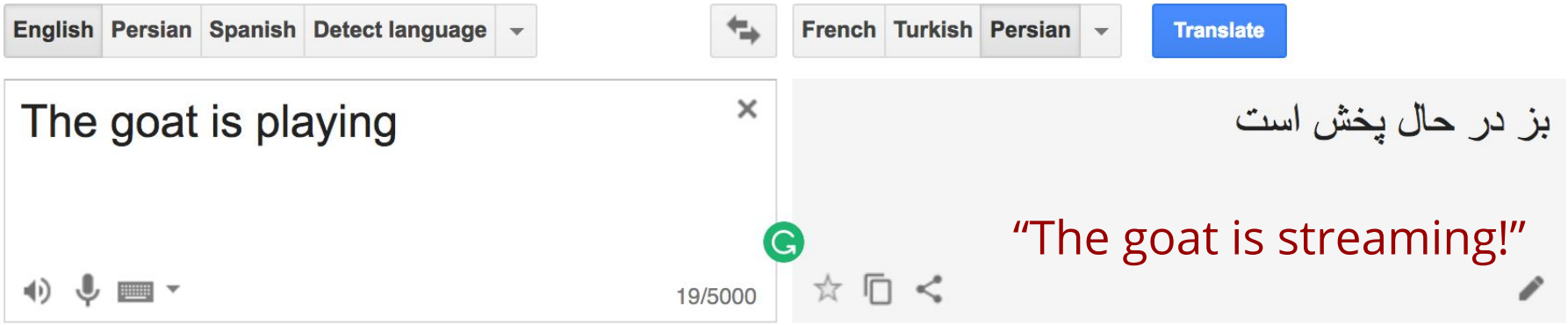
Language Learning in Children

Boz bazi mikonad!

The goat is playing!

Bozbazi mikonad!

Gap Between People and Computers



AI is still far from human performance.

Computational Study of Language

Examine the role of:

- General vs language-specific mechanisms.
- Innate/learned biases.
- Appropriate representations.
- Interaction with other people.

Can we use the insights to improve computers?

Learn, Represent, and Understand

Word learning

- General vs language-specific mechanisms.
- Innate/learned biases.

Comp Sci informs Cog Sci

Word representations

- The appropriate representations.

Cog Sci informs Comp Sci

Understanding mental states

- Interaction with other people.

Learn, Represent, and Understand

Word learning

- **General vs language-specific mechanisms.**
- Innate/learned biases.

Word representations

- The appropriate representations.

Understanding mental states

- Interaction with other people.

Show what is learnable by applying general cognitive mechanisms to word-learning input.

Word Learning

cat ⟷

Cross-situational Learning

People are sensitive to the statistical regularities across situations. [Pinker 1989; Yu & Smith 2007]

A dax!

***dax* means dog**

Look at the dax!

Word Learning Input

[Nematzadeh et al, CogSci 2012]

Input: a sequence of **utterance-scene** pairs.

taken from CHILDES corpora

Construct a **gold-standard lexicon** that provides a ground-truth meaning for each word.

elephant: {SOLID, ANIMAL, ENTITY, ...}

Word Learning Input

Input: a sequence of **utterance-scene** pairs.

Scene represented by sampling features from the meaning of each word in the utterance.

- ➔ **time = 1**
- ➔ **Utterance**: {she, drinks, milk}
- ➔ **Scene**: {ANIMATE, FEMALE, CONSUME, DRINK, ...}
- ➔ **time = 2**
- ➔ **Utterance**: {let's, play, down, here, because, ...}
- ➔ **Scene**: {LET, ..., PLAY, DOWN, DISTANCE, ...}

Word Learning: Formalism

Input: a sequence of utterance-scene pairs.

Utterance: {she, drinks, milk}

Initially all features are equally likely.

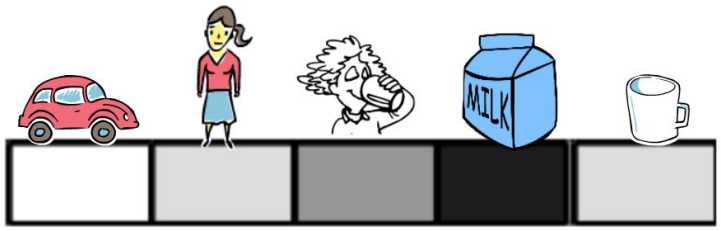
? ? ? ?

Scene: {ANIMATE, FEMALE, CONSUME, DRINK, …}

What features are part of a word's meaning?

Output: word meanings — a probability distribution over a set of features, $P(.|w)$.

milk:



Word Learning: Formalism

[based on Fazly et al. 2010]

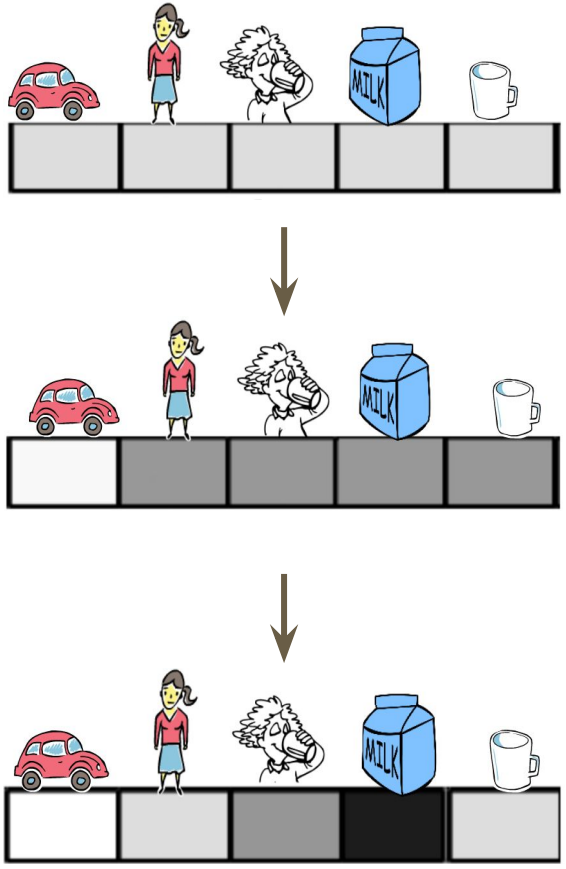
Align words & features using what model knows.

time = 1

Utterance: {she, drinks, milk, ...}

alignment

Scene: {..., FEMALE, ..., IS-LIQUID, ...}



Update what model knows.

time = 2

Utterance: {milk, is, white}

Scene: {IS-LIQUID, ...}

The Model: Foundation

(1) Calculate probability of **alignments** using **learned meanings**.

$$a_t(w, f) = P_t(f|w)$$

association of word and feature

(2) Adjust **learned meanings.**

$$P_{t+1}(f|w) = \frac{\text{assoc}_t(f,\ w)}{\sum_{f_j \in \mathcal{M}} \text{assoc}_t(f_j,\ w)}$$

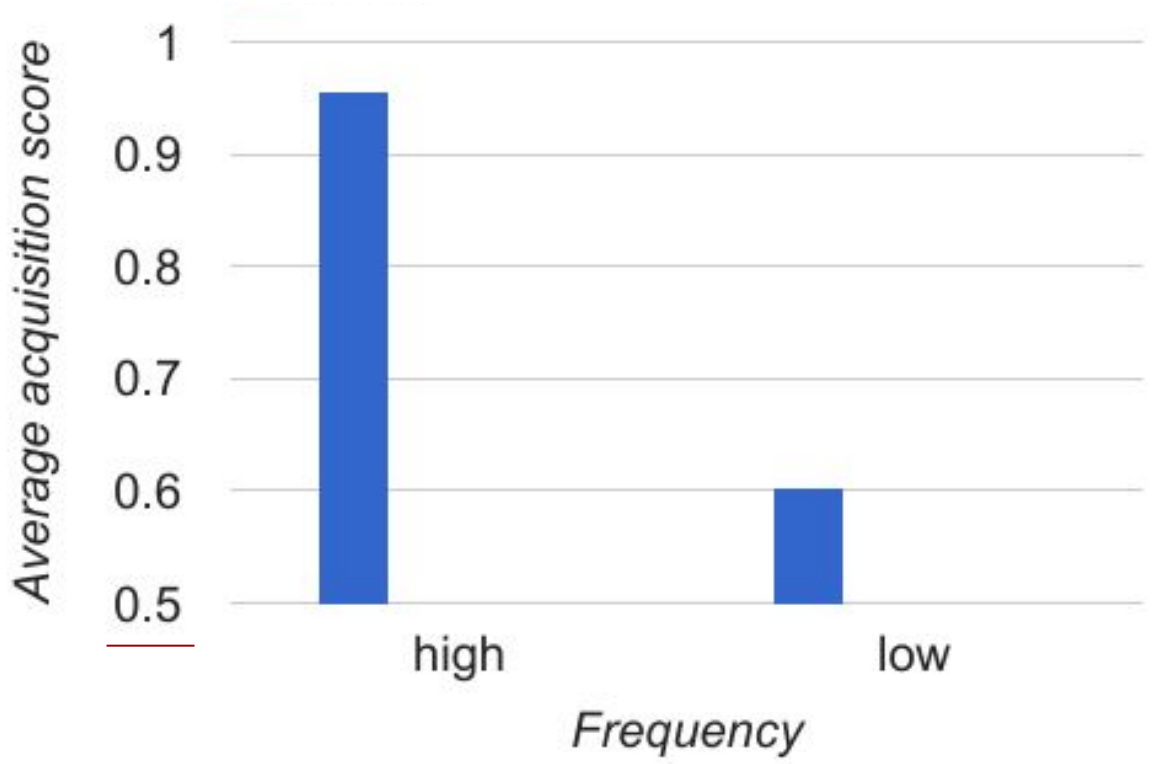
$$\sum_{t'=1}^{t} a_{t'}(w, f)$$

Repeat (1) and (2) for each input pair.

Incremental EM

Word Learning Results

Cosine between learned & gold-standard meaning



The model fails at learning low-frequency words, but children do not.

Learn, Represent, and Understand

Word learning

- General vs language-specific mechanisms.
- **Innate/learned biases.**

Word representations

- The appropriate representations.

Understanding mental states

- Interaction with other people.

Can biases guide learning in more challenging situations?

Mutual Exclusivity Bias in Children

[Markman and Wachtel 1988]

Give me dax

familiar object

unfamiliar object ✓

Limit the number of possible word labels for a familiar object.

Mutual Exclusivity Bias in Model

Limit the number of possible word labels for a familiar object.

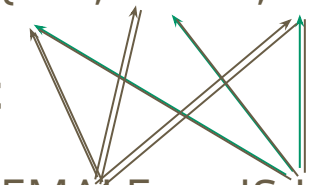
hypotheses

model's knowledge

Add competition (dependence) among words:

Utterance: {she, drinks, milk, ...}

alignment

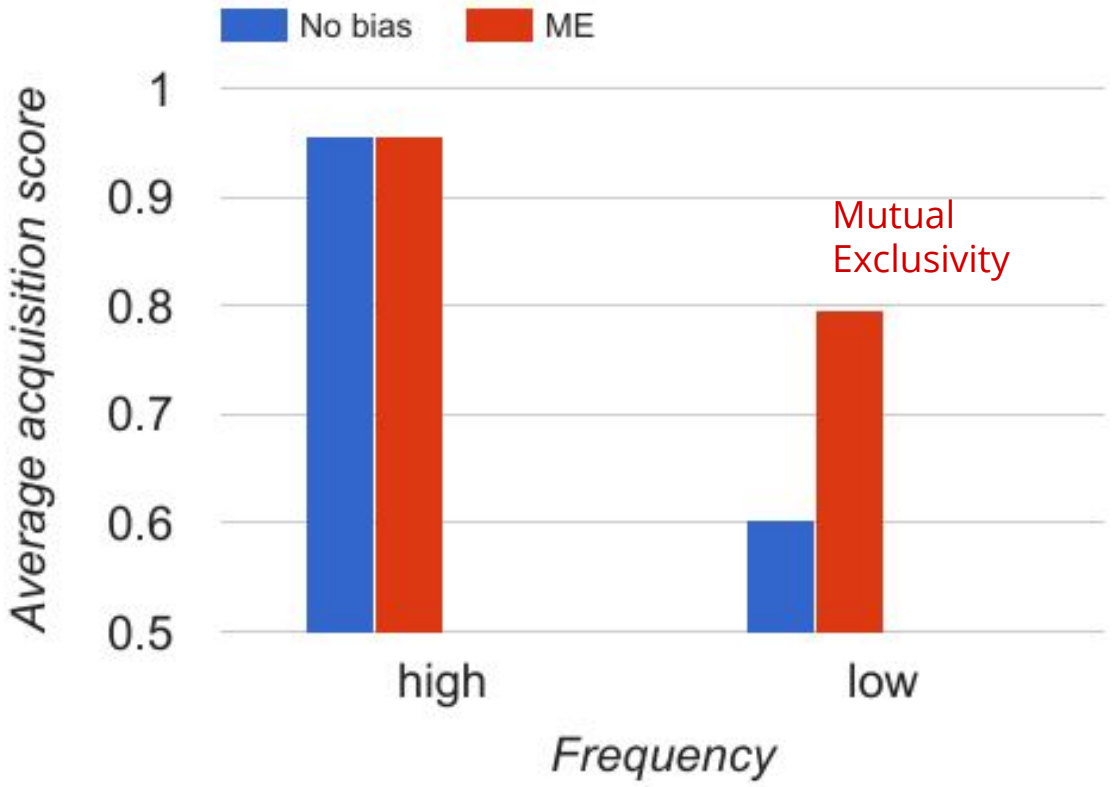


Scene: {..., FEMALE, ..., IS-LIQUID, ...}

$$a_t(w|f) = \frac{P_t(f|w)}{\sum_{w' \in U_t} P_t(f|w')}$$

competition among words

Word Learning Results: ME Bias



The ME bias helps learning low-frequency words, but it is not enough.

The Whole-Object Bias in Children

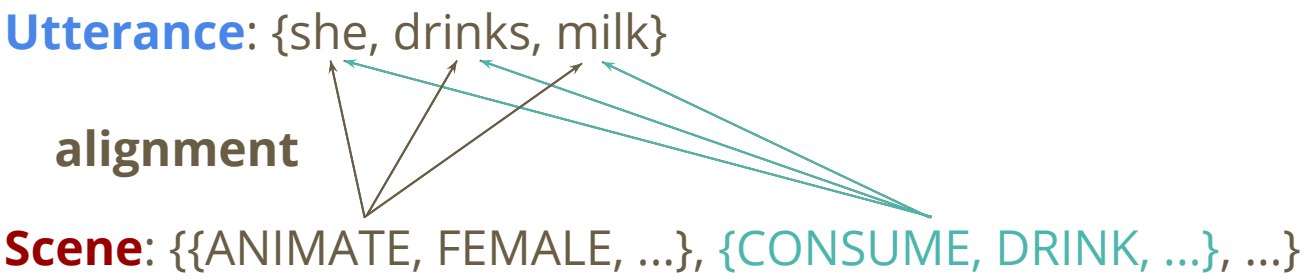
[Markman, 1991]

What is dax?

Learn word labels for the whole object.

The Whole-Object Bias in Model

Represent and align referents as a whole.



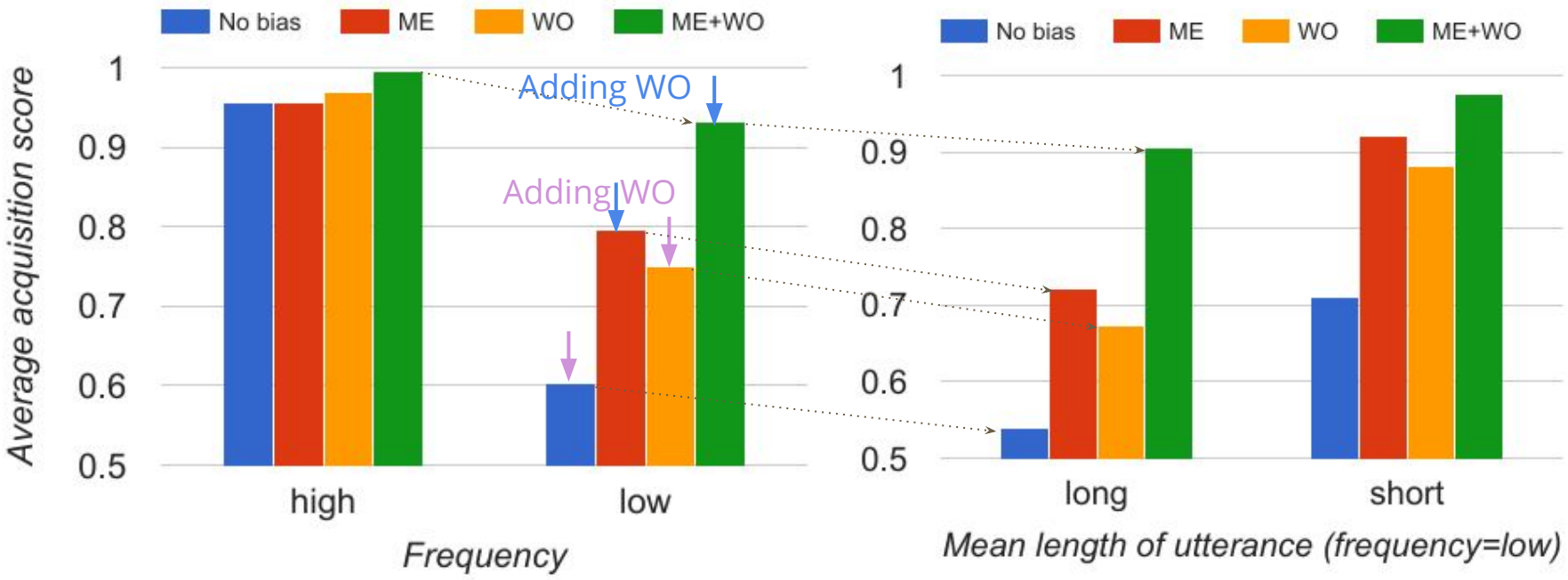
Meaning probability

$$a_t(w,\ r) = \mathrm{sim}(v_t(w),\ v(r)) \qquad a_t(w|r) = \frac{\mathrm{sim}(\mathrm{v}_t(w),\ \mathrm{v}(r))}{\sum_{w' \in \mathrm{U}_t} \mathrm{sim}(\mathrm{v}_t(w'),\ \mathrm{v}(r))}$$

Add ME bias

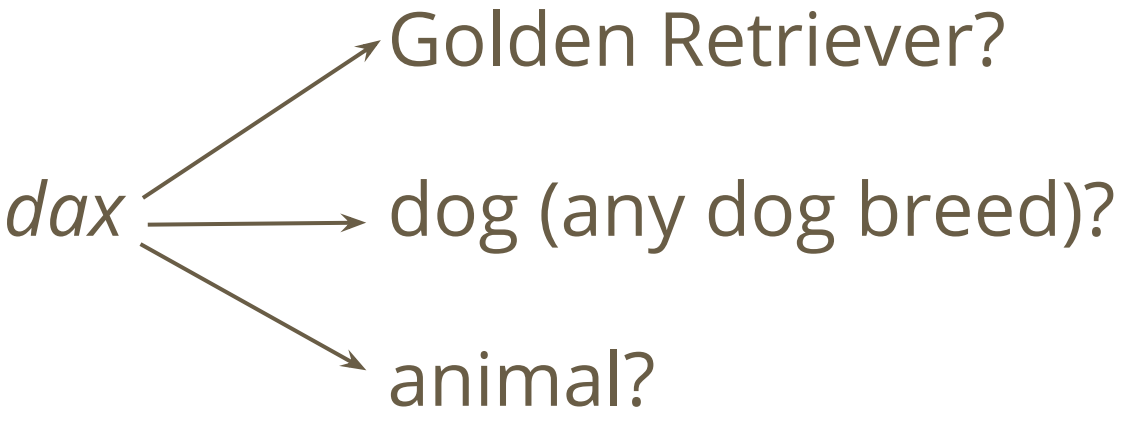
Frequency and Ambiguity

[Nematzadeh, Beekhuizen, Huang, & Stevenson, submitted]



Biases are most important in more challenging learning situations.

Generalizing a Novel Word



Cross-situational statistics are **consistent** with all.

Why dog? A bias that focuses generalization to the **basic-level** (cognitively natural) categories.

Basic-Level Bias in People

[Xu & Tenenbaum, 2007]

Training: (the effect of number of examples)

1-example condition

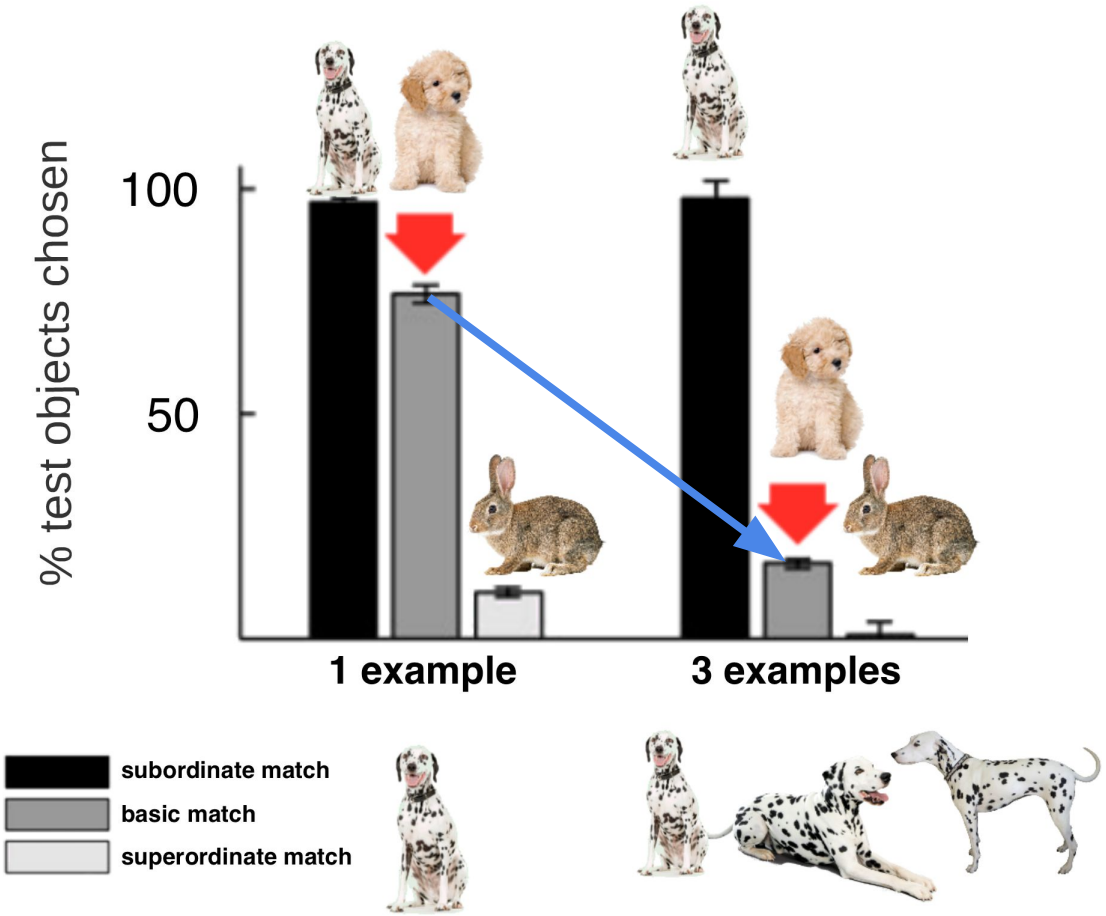
3-example condition

Test: *Pick everything that is a dax.*

Basic-Level Bias in People

[Xu & Tenenbaum, 2007]

basic-level generalization

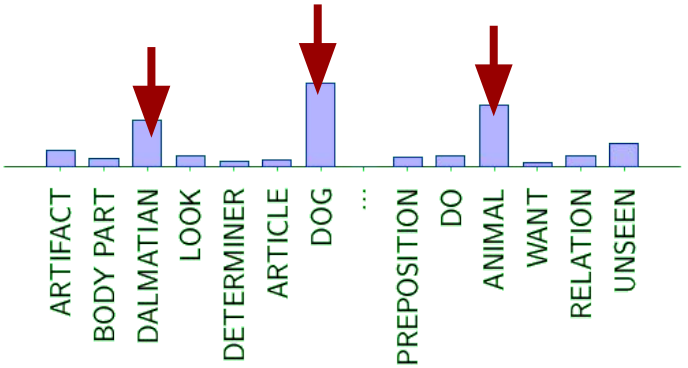


Basic-level generalization is attenuated (**suspicious coincidence**).

Feature Dependencies

Problem: All co-occurring features compete for probability mass.

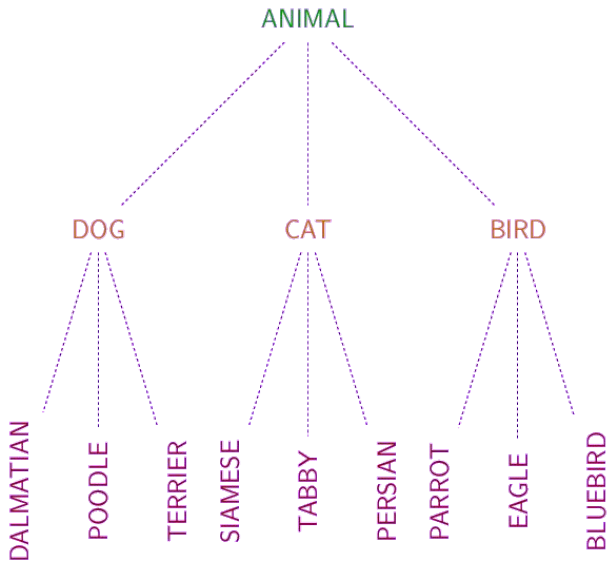
distribution over **all** features



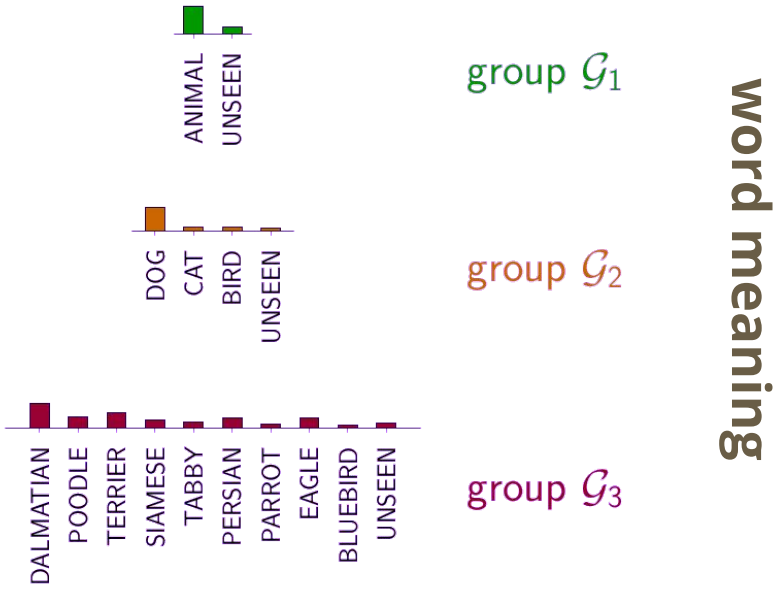
Feature Dependencies

Problem: All co-occurring features compete for probability mass.

Enforce competition among dependent features:



what people know



what the model assumes

Generalization in Model

Associating unobserved features to a word.

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

function of number of types in a group

token frequency of observed features

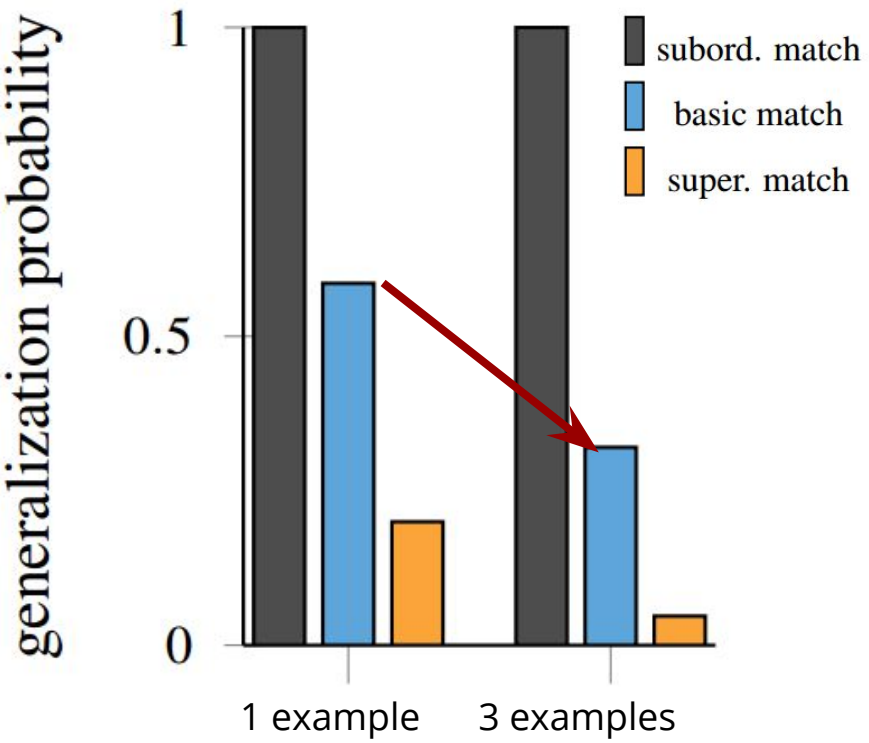
Problem: generalization is only influenced by token frequency.

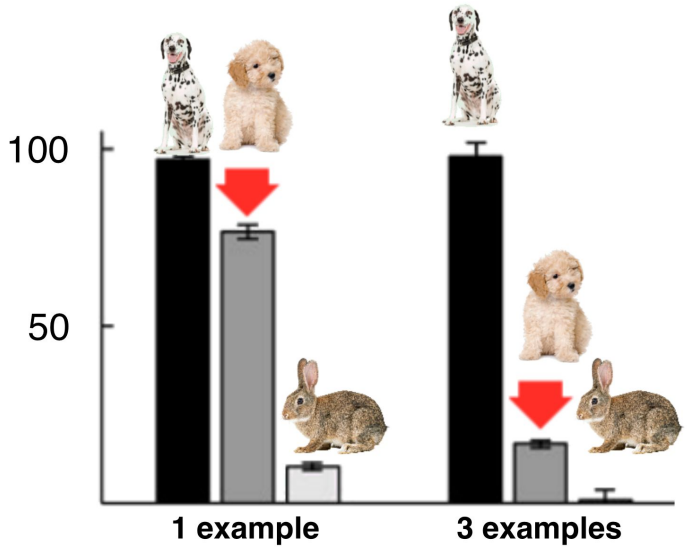
3 types
3 tokens

1 type
3 tokens

Basic-Level Generalization in Model

[Nematzadeh, Grant, & Stevenson, EMNLP 2015]





[Xu & Tenenbaum, 2007]

Replicates the suspicious coincidence effect — decrease in basic-level generalization.

The Role of Biases in Learning

Guide learning in challenging situations:

- Low frequency.
- Ambiguity.
- Identifying the correct-level of generalization.

Can be implemented as:

- Structural assumptions.
- Sensitivity to different statistics.

Learn, Represent, and Understand

Word learning

- General vs language-specific mechanisms.
- Innate/learned biases.

Word representations

- The appropriate representations.

Understanding mental states

- Interaction with other people.

What is the right representation to capture each aspect of human similarity judgements?

Word Representations

milk

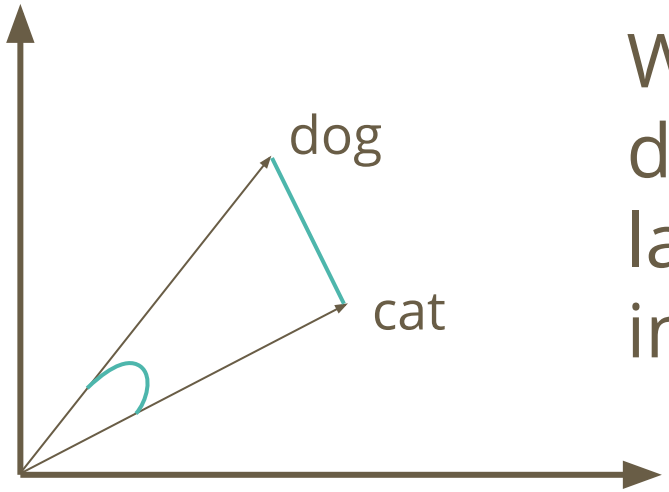
lion

cat

zebra

dog

Vector-Space Representations



Word similarity; word sense disambiguation; semantic role labeling; query expansion; information extraction; ...

cited ~3100 cited ~1200

Recent models (Word2Vec and GloVe)

- Trainable on very large corpora (> 100B words).
- Large coverage (vocabulary size).

Vector-Space Rep: Shortcomings

[Tversky, 1977; Griffiths et al., 2007]

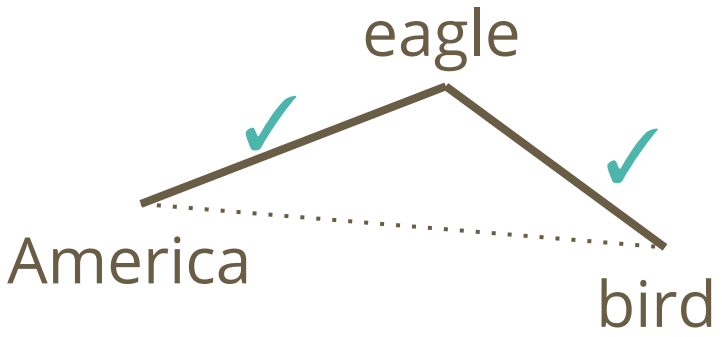
Obey geometric constraints of Euclidean spaces:

- Asymmetry in similarity judgements.

text → book → read

cobra ⇢ snake

- Triangle Inequality.



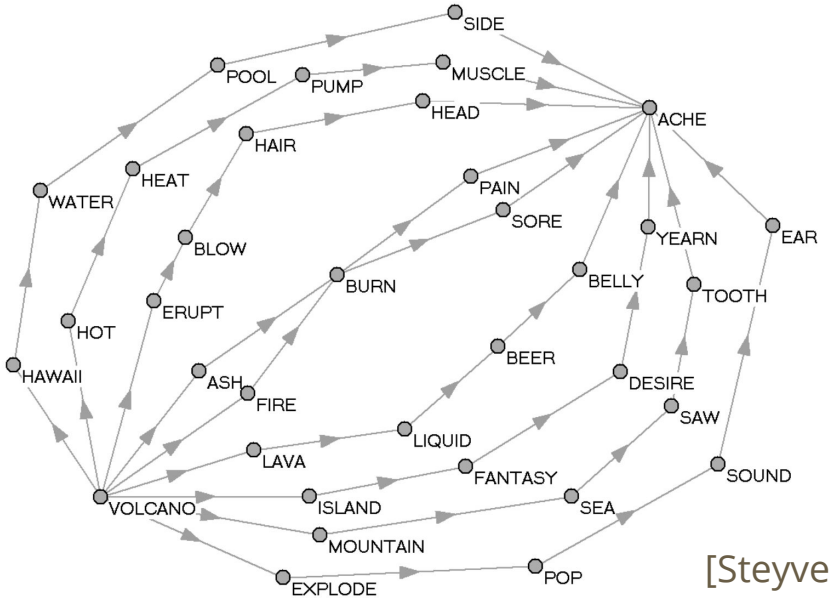
Do the recent models suffer from this?

Evaluating Vector-Space Rep.

[Nematzadeh, Meylan, & Griffiths, 2017]

Data: Nelson association norms. [Nelson et *al.*, 1998]

- Table: chair (0.77), cloth (0.03), eat (0.03)

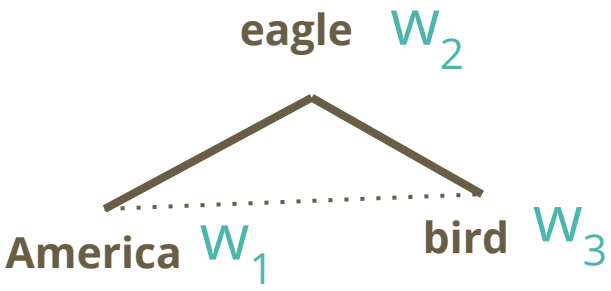


[Steyvers & Tenenbaum 2005]

Measure: conditional probability (not distance).

Results: Triangle Inequality

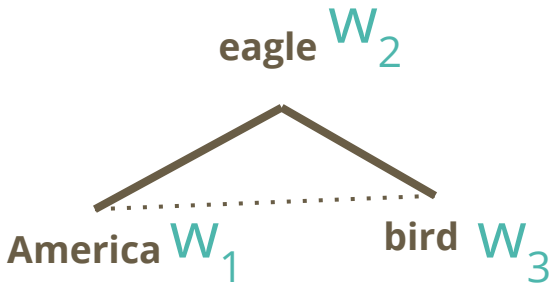
Select tuples where both $P(w_2|w_1)$ and $P(w_3|w_2)$ are greater than a threshold.



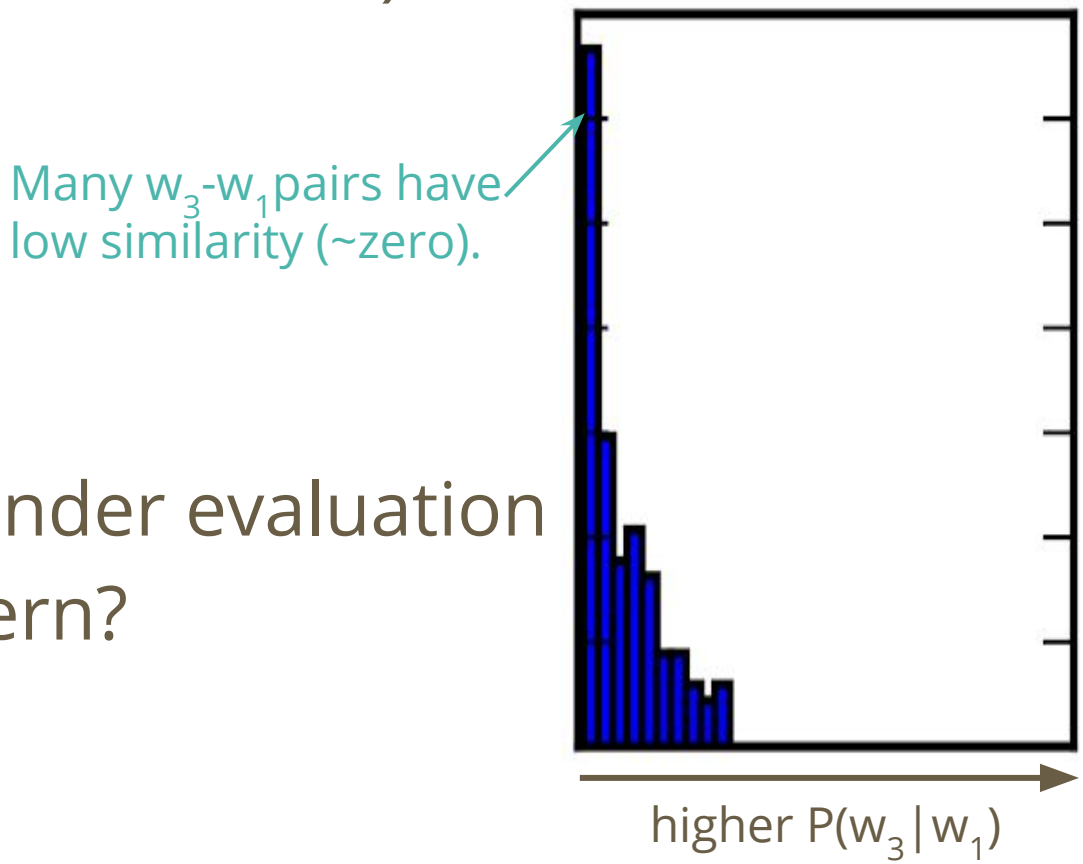
Plot distribution of $P(w_3|w_1)$.

How does the distribution look for Nelson norms?

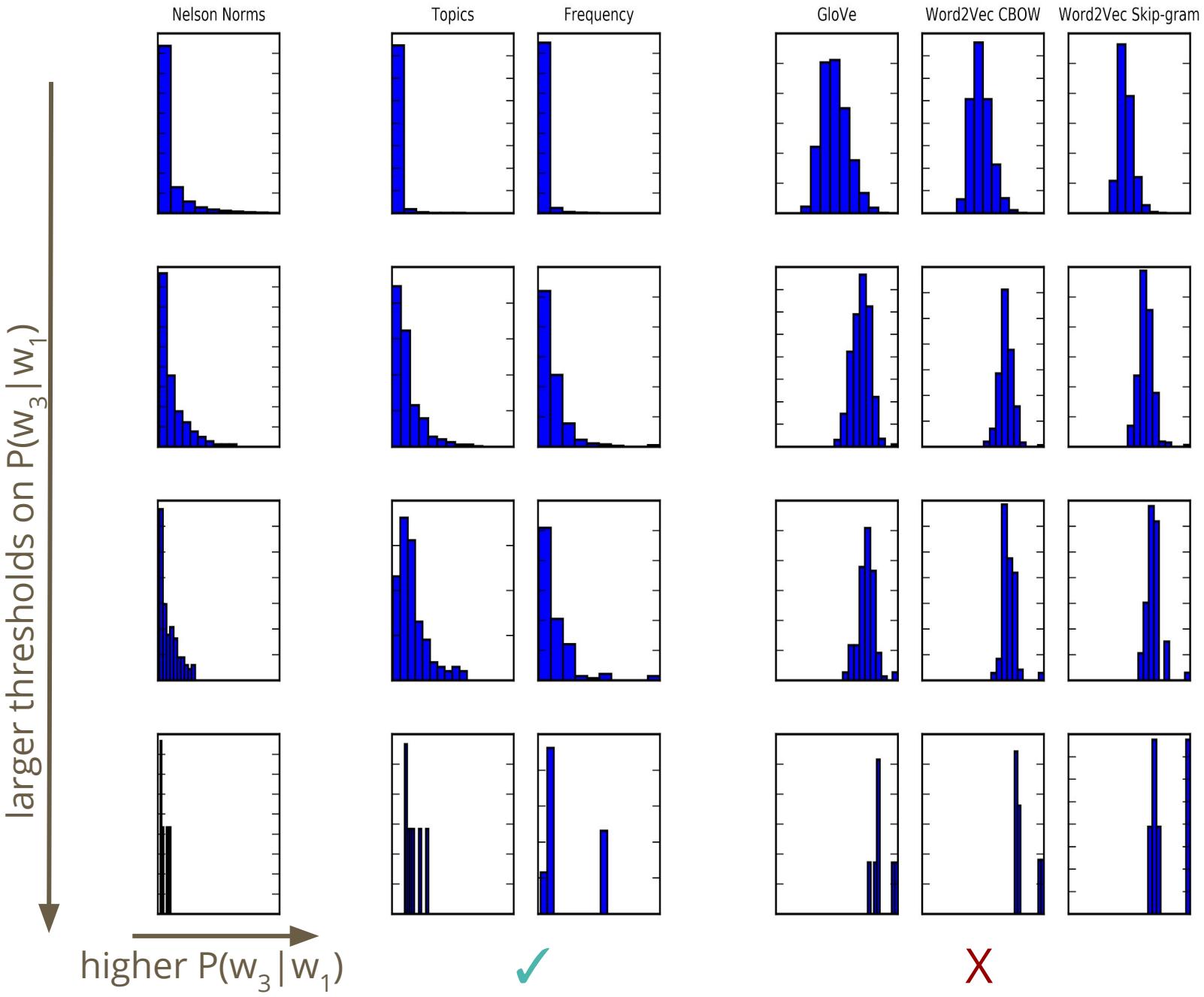
Results: Triangle Inequality



Distribution of $P(w_3|w_1)$ in Nelson norms (for a high threshold):



Do the models under evaluation predict this pattern?


Nelson Norms
Topics
Frequency
GloVe
Word2Vec CBOW
Word2Vec Skip-gram
larger thresholds on P(w3 | w1)
higher P(w3 | w1)
✓
X

Summary of Results

Recent vector-space representations are good at capturing overall associations given very large corpora.

But cannot predict aspects of human semantic processing -- when the observed patterns do not obey the constraints of vector-spaces.

Consider representations that more directly capture word-pair probabilities.

Learn, Represent, and Understand

Word learning

- General vs language-specific mechanisms.
- Innate/learned biases.

Word representations

- The appropriate representations.

Understanding mental states

- Interaction with other people.

Can causal structure of interactions help models understand beliefs?

Understanding Mental States

Correctly Answering Questions

Require reasoning (not simply memorizing).

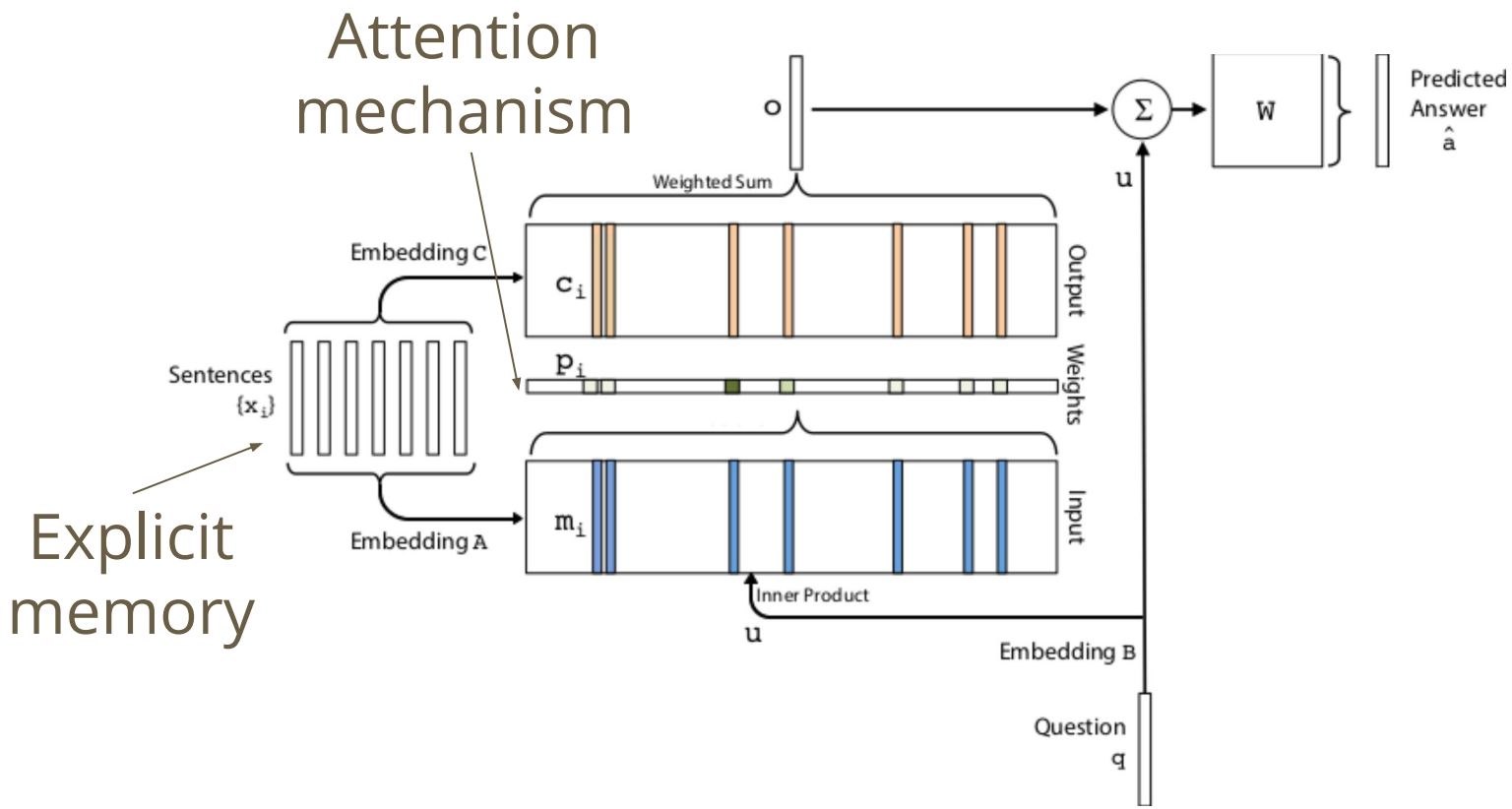
Mary got the milk there.
John moved to the bedroom.
Sandra went back to the kitchen.
Mary travelled to the hallway.

Q: *Where is the milk?* ***A:*** *hallway*

Evaluate this capacity using Facebook bAbi dataset. [Weston et al., 2016]

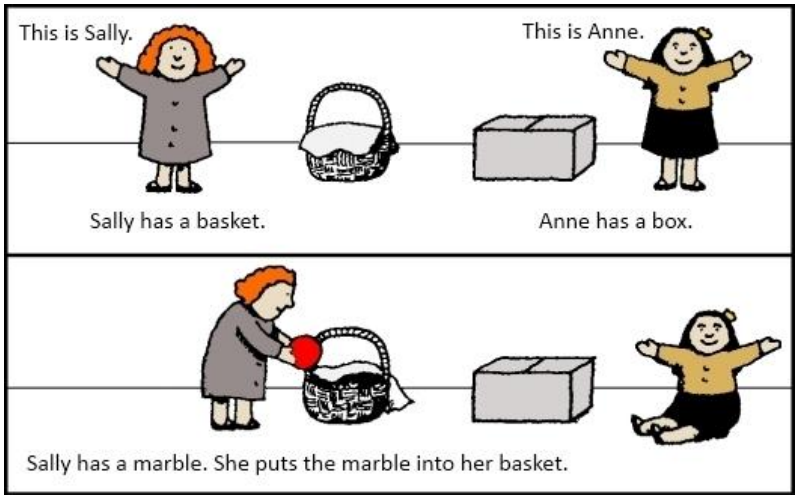
End-to-End Memory Network

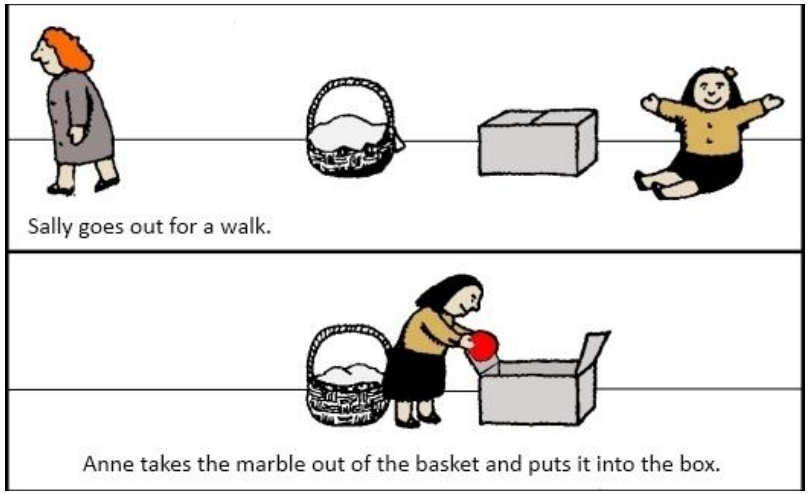
[Sukhbaatar et *al*., 2015]



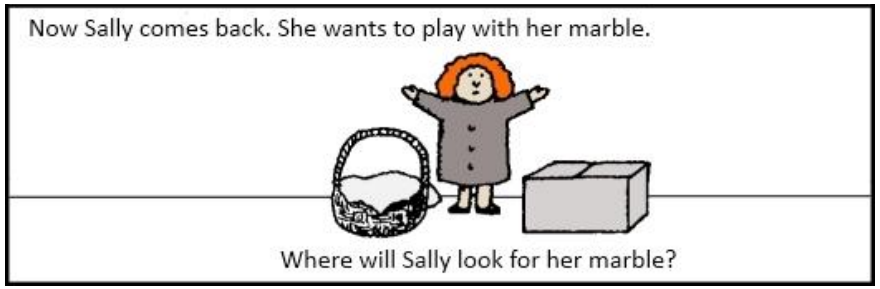
The best model fails at only 4 out of 20 bAbi tasks.

Reasoning about Actions & Beliefs





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



Need to reason about others' beliefs & actions

Does memory network succeed in such tasks?

Belief → Action and Action→ Belief

True Belief

Sally believes the milk is in the pantry.
Anne moved the milk to the fridge.
Q: Where did Sally search for the milk?
A: Fridge.

True Belief

Sally placed the milk in the pantry.
Anne moved the milk to the fridge.
Q: Where does Sally believe the milk is?
A: Fridge.

False Belief

Sally believes the milk is in the pantry.
Sally exited the kitchen.
Anne moved the milk to the fridge.
Sally entered the kitchen.
Q: Where did Sally search for the milk?
A: Pantry.

False Belief

Sally placed the milk in the pantry.
Sally exited the kitchen.
Anne moved the milk to the fridge.
Sally entered the kitchen.
Q: Where does Sally believe the milk is?
A: Pantry.

Action→ Belief → Action

True Belief

Sally placed the milk in the pantry.
Anne moved the milk to the fridge.
Q: Where did Sally search for the milk?
A: Fridge.

False Belief

Sally placed the milk in the pantry.
Sally exited the kitchen.
Anne moved the milk to the fridge.
Sally entered the kitchen.
Q: Where did Sally search for the milk?
A: Pantry.

Sally-Anne or False-belief Task

Evaluating the Memory Network

[Grant, Nematzadeh, & Griffiths, submitted]

Train the model on various conditions:

- B→ A
- A→ B
- A→ B→ A ✓
- A→ B + B→ A (transitive inference)
- A→ B + B→ A + A→ B→ A

Performs ok when the training & test has the same structure.

Test the model:

- A→ B→ A (Sally-Anne task)

Modeling Participants' Beliefs

Extend the model with:

- separate memories for Sally, Anne & observer,
- ability to attend to each memory.

Performance increases in all tasks; significantly when trained on A→ B **+** B→ A **+** A→ B→ A.

Need to model each participant's mental state to correctly answer questions.

Learn, Represent, and Understand

Word learning

- General vs language-specific mechanisms.
- Innate/learned biases.

Word representations

- The appropriate representations.

Understanding mental states

- Interaction with other people.

Future Directions

How do we learn a language?
How can this inform AI systems?

Deep Representations vs. Humans

Deep nets are good at representation learning.

Typicality ratings. [Lake et al., 2015]

Similarity judgements. [Peterson et al., 2016]

Do deep representations replicate human language performance?

Deep Representations of Language

Predicting a formal language:

- context-free [Wiles and Elman, 1995]
- context-sensitive [Gers and Schmidhuber, 2001]

Modeling compositionality.

- Representations for roles (in addition to words).
- $AB = Axr_1 + Bxr_2$ [Smolensky, 2006]

Representations in Seq2Seq models:

- capturing syntax and semantics.

Attention in Children & Computers

Children only attend to aspects of environment.

- Helps learning.
- Reduces complexity.

Recent neural nets also use attention.

- Helps learning.
- Increases complexity.

Attention in Children & Computers

How similar is attention in neural nets & children?

Compare neural attention mechanisms with eye-tracking data.

Can attention in neural models simplify learning?

Language in People and Computers

Human language learning is a complex process.

AI systems need to address similar challenges to those people face in language learning.

A multi-disciplinary study of language benefits both AI and cognitive science.

- Comp Sci informs Cog Sci.
- Cog Sci informs Comp Sci.

Other Projects & Acknowledgments

Desirable difficulties in learning.
[Nematzadeh et al, 2012, 2013, submitted]

Individual differences in word learning. [Nematzadeh et al, 2011, 2012, 2014]

Learning multiword expressions.
[Fazly et al, 2009; Nematzadeh et al, 2013]

Learning, organizing, and searching semantic knowledge.
[Nematzadeh et al 2015, 2016; Miscevic et al, submitted]

Barend Beekhuizen
Afsaneh Fazly
Erin Grant
Tom Griffiths
Shanshan Huang
Stephan Meylan
Filip Miscevic
Suzanne Stevenson

Thanks!