Learnability and Semantic Universals

Shane Steinert-Threlkeld
(joint work-in-progress with Jakub Szymanik)

Cognitive Semantics and Quantities Kick-off Workshop
September 28, 2017

Overview

Universals in Linguistic Theory

Question

What is the range of variation in human languages? That is: which out of all of the logically possible languages that humans could speak, do they in fact speak?

Example Universals

- Sound (phonology):
  All languages have consonants and vowels. Every language has at least /a i u/.

Example Universals

- Sound (phonology):
  All languages have consonants and vowels. Every language has at least /a i u/.
- Grammar (syntax):
  All languages have verbs and nouns.

Example Universals

- Sound (phonology):
  All languages have consonants and vowels. Every language has at least /a i u/.
- Grammar (syntax):
  All languages have verbs and nouns.
- Meaning (semantics):

Example Universals

- Sound (phonology):
  All languages have consonants and vowels. Every language has at least /a i u/.
- Grammar (syntax):
  All languages have verbs and nouns.
- Meaning (semantics):
  All languages have syntactic constituents (NPs) whose semantic function is to express generalized quantifiers. (Barwise and Cooper 1981)

Determiners

- Determiners:
  - Simple: *every*, *some*, *few*, *most*, *five*, ...
  - Complex: *all but five*, *fewer than three*, *at least eight or fewer than five*, ...

Determiners

- Determiners:
  - Simple: *every*, *some*, *few*, *most*, *five*, ...
  - Complex: *all but five*, *fewer than three*, *at least eight or fewer than five*, ...
- Denote type $\langle 1, 1 \rangle$ generalized quantifiers: sets of models of the form $\langle M, A, B \rangle$ with $A, B \subseteq M$

Determiners

- Determiners:
  - Simple: *every*, *some*, *few*, *most*, *five*, ...
  - Complex: *all but five*, *fewer than three*, *at least eight or fewer than five*, ...
- Denote type $\langle 1, 1 \rangle$ generalized quantifiers: sets of models of the form $\langle M, A, B \rangle$ with $A, B \subseteq M$
- For example:

$$[\![every]\!] = \{\langle M, A, B \rangle : A \subseteq B\}$$
$$[\![most]\!] = \{\langle M, A, B \rangle : |A \cap B| > |A \setminus B|\}$$

Monotonicity

- Many French people smoke cigarettes
  $\Rightarrow$ Many French people smoke

Monotonicity

- Many French people smoke cigarettes
  $\Rightarrow$ Many French people smoke
- Few French people smoke
  $\Rightarrow$ Few French people smoke cigarettes

Monotonicity

- Many French people smoke cigarettes
  $\Rightarrow$ Many French people smoke
- Few French people smoke
  $\Rightarrow$ Few French people smoke cigarettes
- Q is *upward (resp. downward) monotone* iff:
  if $\langle M, A, B \rangle \in$ Q and $B \subseteq B'$ (resp. $B' \subseteq B$), then $\langle M, A, B \rangle \in$ Q
- A determiner is *monotone* if it denotes either an upward or downward monotone generalized quantifier

Monotonicity Universal

Monotonicity Universal (Barwise and Cooper 1981)

All simple determiners are monotone.

Quantity

- Key intuition: determiners denote *general* relations between their restrictor and nuclear scope, not ones that depend on the identities of particular elements of the domain

Quantity

- Key intuition: determiners denote *general* relations between their restrictor and nuclear scope, not ones that depend on the identities of particular elements of the domain
- Q is *permutation-closed* iff:
  if $\langle M, A, B \rangle \in \mathrm{Q}$ and π is a permutation of M, then
  $\pi(\langle M, A, B \rangle) \in \mathrm{Q}$

Quantity

- Key intuition: determiners denote *general* relations between their restrictor and nuclear scope, not ones that depend on the identities of particular elements of the domain
- Q is *permutation-closed* iff:
  if $\langle M, A, B \rangle \in \mathrm{Q}$ and π is a permutation of M, then $\pi(\langle M, A, B \rangle) \in \mathrm{Q}$
- Q is *isomorphism-closed* iff:
  if $\langle M, A, B \rangle \in \mathrm{Q}$ and $\langle M, A, B \rangle \cong \langle M', A', B' \rangle$, then $\langle M', A', B' \rangle \in \mathrm{Q}$

Quantity

- Key intuition: determiners denote *general* relations between their restrictor and nuclear scope, not ones that depend on the identities of particular elements of the domain
- Q is *permutation-closed* iff:
  if $\langle M, A, B\rangle \in \mathrm{Q}$ and π is a permutation of M, then $\pi(\langle M, A, B\rangle) \in \mathrm{Q}$
- Q is *isomorphism-closed* iff:
  if $\langle M, A, B\rangle \in \mathrm{Q}$ and $\langle M, A, B\rangle \cong \langle M', A', B'\rangle$, then $\langle M', A', B'\rangle \in \mathrm{Q}$
- Q is *quantitative*:
  if $\langle M, A, B\rangle \in \mathrm{Q}$ and $A \cap B, A \setminus B, B \setminus A, M \setminus (A \cup B)$ have the same cardinality as their primed-counterparts, then $\langle M', A', B'\rangle \in \mathrm{Q}$

Quantity

- Keenan and Stavi 1986, p. 311:
  "Monomorphemic dets are logical [= permutation-closed]."

Quantity

- Keenan and Stavi 1986, p. 311:
  "Monomorphemic dets are logical [= permutation-closed]."
- Peters and Westerståhl 2006, p. 330:
  "All lexical quantifier expressions in natural languages denote ISOM [= isomorphism-closed] quantifiers."

Quantity Universal

Quantity Universal

All simple determiners are quantitative.

Conservativity

- Key intuition: the restrictor *restricts* what the determiner talks about

Conservativity

- Key intuition: the restrictor *restricts* what the determiner talks about
- Many French people smoke cigarettes
  $\equiv$ Many French people are French people who smoke cigarettes

Conservativity

- Key intuition: the restrictor *restricts* what the determiner talks about
- Many French people smoke cigarettes
  $\equiv$ Many French people are French people who smoke cigarettes
- Q is *conservative* iff:
  $\langle M, A, B \rangle \in \mathrm{Q}$ if and only if $\langle M, A, A \cap B \rangle \in \mathrm{Q}$

Conservativity Universal

Conservativity Universal (Barwise and Cooper 1981)

All simple determiners are conservative.

Explaining Universals

Natural Question

Why do the universals hold? What explains the limited range of quantifiers expressed by simple determiners?

Explaining Universals

Natural Question

Why do the universals hold? What explains the limited range of quantifiers expressed by simple determiners?

- Answer 1: *learnability*.
  (Barwise and Cooper 1981; Keenan and Stavi 1986; Szabolcsi 2010)

Explaining Universals

Natural Question

Why do the universals hold? What explains the limited range of quantifiers expressed by simple determiners?

- Answer 1: *learnability*.
  (Barwise and Cooper 1981; Keenan and Stavi 1986; Szabolcsi 2010)
- The universals greatly restrict the search space that a language learner must explore when learning the meanings of determiners. This makes it easier (possible?) for them to learn such meanings from relatively small input.
  [Compare: Poverty of the Stimulus argument for UG.]

Explaining Universals

Natural Question

Why do the universals hold? What explains the limited range of quantifiers expressed by simple determiners?

- Answer 1: *learnability*.
  (Barwise and Cooper 1981; Keenan and Stavi 1986; Szabolcsi 2010)
- The universals greatly restrict the search space that a language learner must explore when learning the meanings of determiners. This makes it easier (possible?) for them to learn such meanings from relatively small input.
  [Compare: Poverty of the Stimulus argument for UG.]
- In a sense must be true, but:
  "Likely, the unrestricted space has many hypotheses which are so implausible, they can be ignored quickly and do not affect learning. The hard part of learning, may be choosing between the plausible competitor meanings, not in weeding out a large space of potential meanings." Piantadosi 2013, p. 22

Explaining Universals

Natural Question

Why do the universals hold? What explains the limited range of quantifiers expressed by simple determiners?

- Answer 2: *learnability*.
  (Peters and Westerståhl 2006)

Explaining Universals

Natural Question

Why do the universals hold? What explains the limited range of quantifiers expressed by simple determiners?

- Answer 2: *learnability*.
  (Peters and Westerståhl 2006)
- The universals aid learnability because quantifiers satisfying the universals are *easier* to learn than those that do not.

Explaining Universals

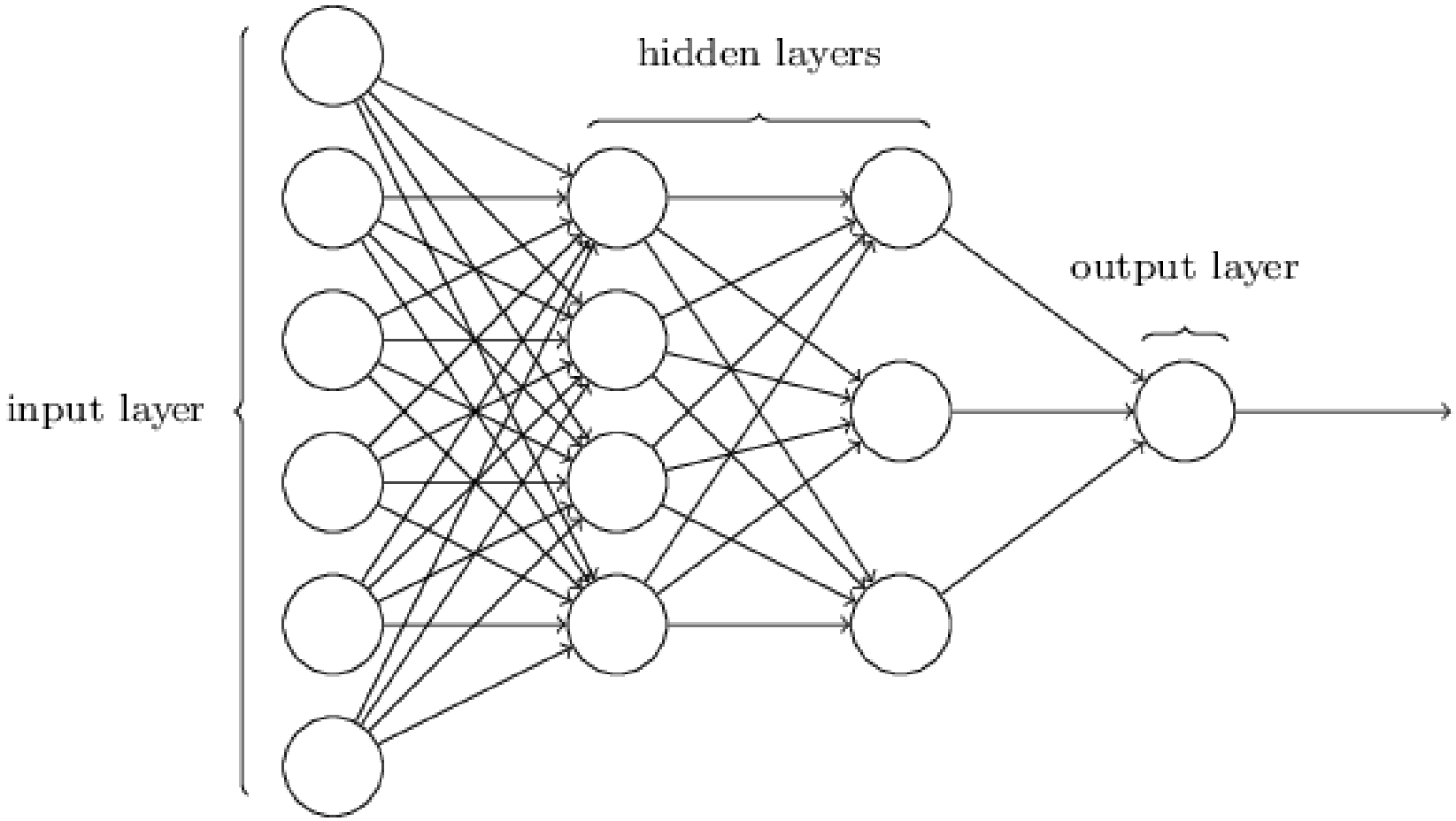
Natural Question

Why do the universals hold? What explains the limited range of quantifiers expressed by simple determiners?

- Answer 2: *learnability*.
  (Peters and Westerståhl 2006)
- The universals aid learnability because quantifiers satisfying the universals are *easier* to learn than those that do not.
- Challenge: provide a model of learning which makes good on this promise.
  [See Tiede 1999; Gierasimczuk 2007; Gierasimczuk 2009 for earlier attempts.]

Overview

Neural Networks



Source: Nielsen, "Neural Networks and Deep Learning", Determination Press

Gradient Descent and Back-propogation

The learning framework will be a *non-convex optimization* problem.

- A total loss function, which will be the mean of a 'local' error function:

$$
\begin{aligned}
L &= \frac{1}{N} \sum \ell(\hat{y}_i, y_i) \\
&= \frac{1}{N} \sum \ell(\mathrm{NN}(\vec{\theta}, x_i), y_i)
\end{aligned}
$$

Gradient Descent and Back-propogation

The learning framework will be a *non-convex optimization* problem.

- A total loss function, which will be the mean of a 'local' error function:

$$
\begin{aligned}
L &= \frac{1}{N} \sum \ell(\hat{y}_i, y_i) \\
&= \frac{1}{N} \sum \ell(\mathrm{NN}(\vec{\theta}, x_i), y_i)
\end{aligned}
$$

- Navigate the landscape of weight space towards lower loss:

$$\vec{\theta}_{t+1} \leftarrow \vec{\theta}_t - \alpha \nabla_{\vec{\theta}} L$$

Gradient Descent and Back-propogation

The learning framework will be a *non-convex optimization* problem.

- A total loss function, which will be the mean of a 'local' error function:

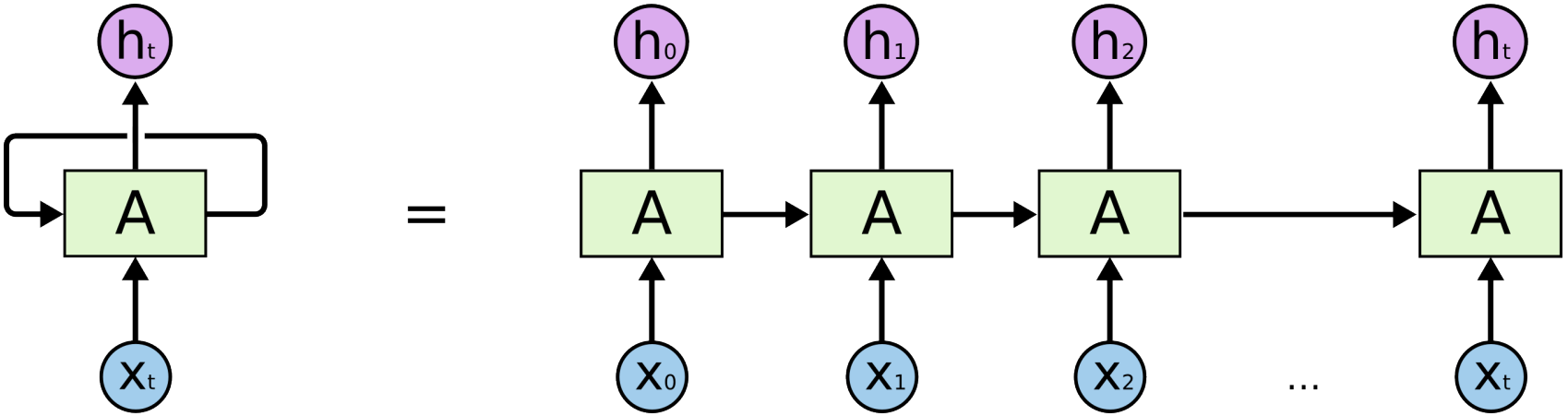
$$L = \frac{1}{N} \sum \ell(\hat{y}_i, y_i)$$
$$= \frac{1}{N} \sum \ell(\mathrm{NN}(\vec{\theta}, x_i), y_i)$$

- Navigate the landscape of weight space towards lower loss:

$$\vec{\theta}_{t+1} \leftarrow \vec{\theta}_t - \alpha \nabla_{\vec{\theta}} L$$

- Back-propagation: calculate $\ell(\cdot, \cdot)$ after a forward-pass of the network. 'Propagate' the error backwards through the network to calculate the gradient.

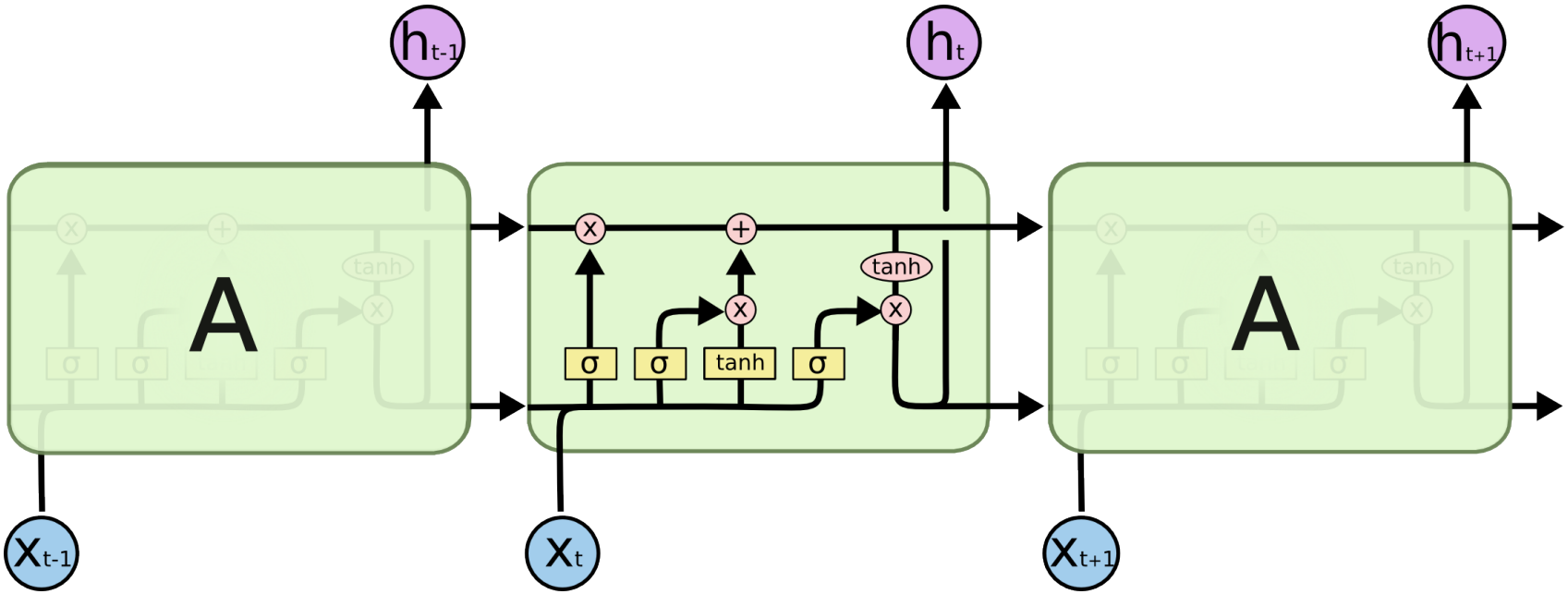
Recurrent Neural Networks



Source: Olah, "Understanding LSTM Networks", http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Long Short-Term Memory

Introduced in Hochreiter and Schmidhuber 1997 to solve the exploding/vanishing gradients problem.



Source: Olah, "Understanding LSTM Networks", http://colah.github.io/posts/2015-08-Understanding-LSTMs/

The Learning Task

- Our data will be the following:
  - Input: $\langle \mathsf{Q}, \mathcal{M} \rangle$ pairs, presented sequentially
  - Output: T/F, depending on whether $\mathcal{M} \in \mathsf{Q}$

The Learning Task

- Our data will be the following:
  - Input: $\langle \mathsf{Q}, \mathcal{M} \rangle$ pairs, presented sequentially
  - Output: T/F, depending on whether $\mathcal{M} \in \mathsf{Q}$
- So, this will be a *sequence classification* task.

The Learning Task

- Our data will be the following:
  - Input: $\langle \mathrm{Q}, \mathcal{M} \rangle$ pairs, presented sequentially
  - Output: T/F, depending on whether $\mathcal{M} \in \mathrm{Q}$
- So, this will be a *sequence classification* task.
- The loss function we minimize is *cross-entropy*. In our case, with $y \in \{0, 1\}$, very simple:

$$\ell(\mathrm{NN}(\vec{\theta}, x), y) = -\ln(\mathrm{NN}(\vec{\theta}, x)_y)$$

Data Generation

Algorithm 1 Data Generation Algorithm

Inputs: max_len, num_data, quants

```
data ← []
while len(data) < num_data do
    N ∼ Unif([max_len])
    Q ∼ Unif(quants)
    cur_seq ← Unif({A ∩ B, A \ B, B \ A, M \ (A ∪ B)}, N)
    if ⟨Q, cur_seq⟩ ∉ data then
        data.append(generate_point(⟨Q, cur_seq, Q ∈ cur_seq?⟩))
    end if
end while
return shuffle(data)
```

Overview

General Set-up

For each proposed semantic universal:

General Set-up

For each proposed semantic universal:

- Choose a set of minimally different quantifiers that disagree with respect to the universal

General Set-up

For each proposed semantic universal:

- Choose a set of minimally different quantifiers that disagree with respect to the universal
- Run some number of trials of training an LSTM to learn those quantifiers

General Set-up

For each proposed semantic universal:

- Choose a set of minimally different quantifiers that disagree with respect to the universal
- Run some number of trials of training an LSTM to learn those quantifiers
- Stop when:
  500 minibatch running mean total accuracy on test set is > 0.99; or
  mean probability assigned to correct truth-value on test set is > 0.99

General Set-up

For each proposed semantic universal:

- Choose a set of minimally different quantifiers that disagree with respect to the universal
- Run some number of trials of training an LSTM to learn those quantifiers
- Stop when:
  500 minibatch running mean total accuracy on test set is > 0.99; or
  mean probability assigned to correct truth-value on test set is > 0.99
- Measure, for each quantifier, its *convergence point*:
  first i such that mean accuracy from i onward for Q is > 0.98

General Set-up

For each proposed semantic universal:

- Choose a set of minimally different quantifiers that disagree with respect to the universal
- Run some number of trials of training an LSTM to learn those quantifiers
- Stop when:
  500 minibatch running mean total accuracy on test set is > 0.99; or
  mean probability assigned to correct truth-value on test set is > 0.99
- Measure, for each quantifier, its *convergence point*:
  first i such that mean accuracy from i onward for Q is > 0.98

Code and Data Available At:

http://github.com/shanest/quantifier-rnn-learning

Experiment 1: Monotonicity

Monotonicity Universal (Barwise and Cooper 1981)

All simple determiners are monotone.

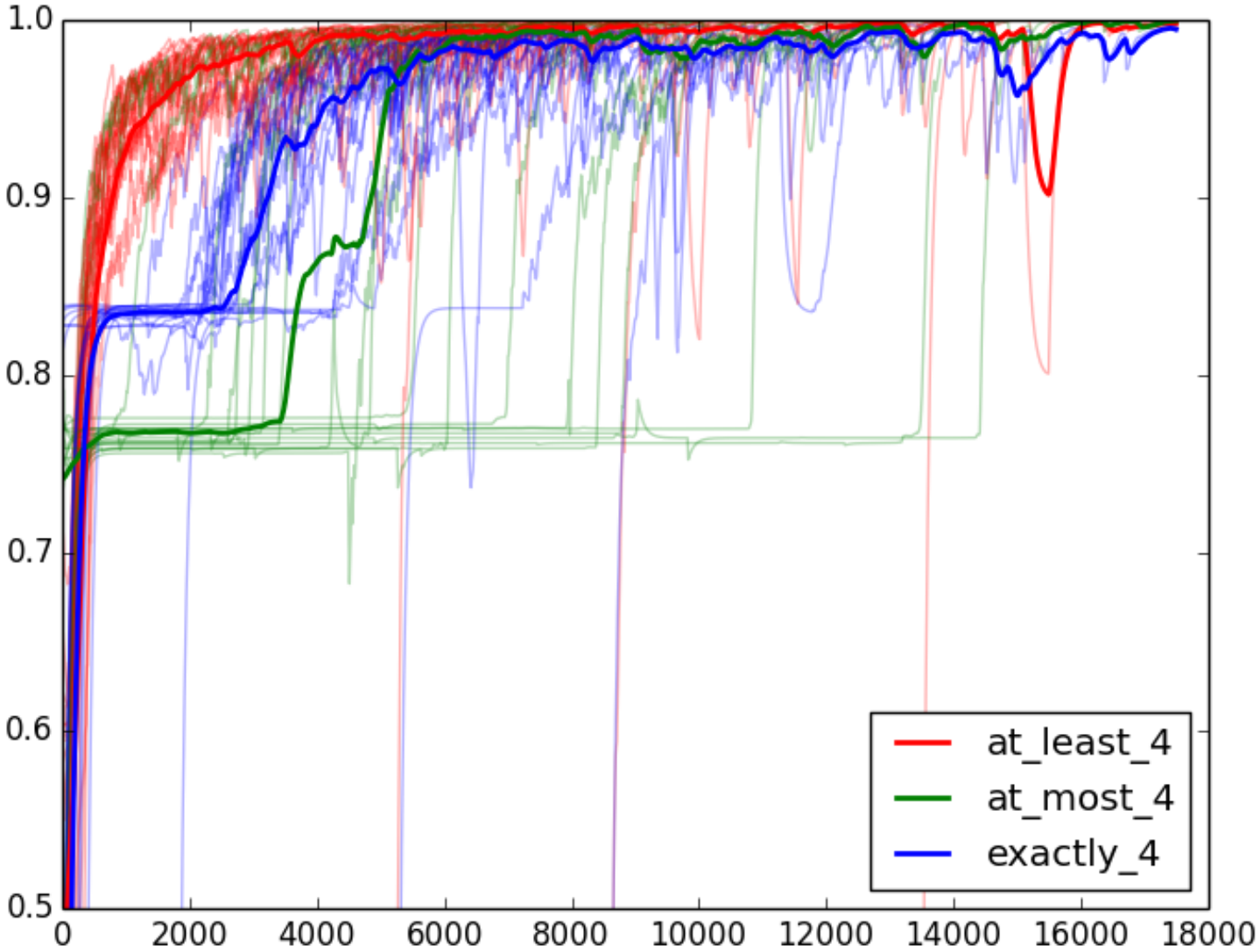
Experiment 1: Monotonicity

Monotonicity Universal (Barwise and Cooper 1981)

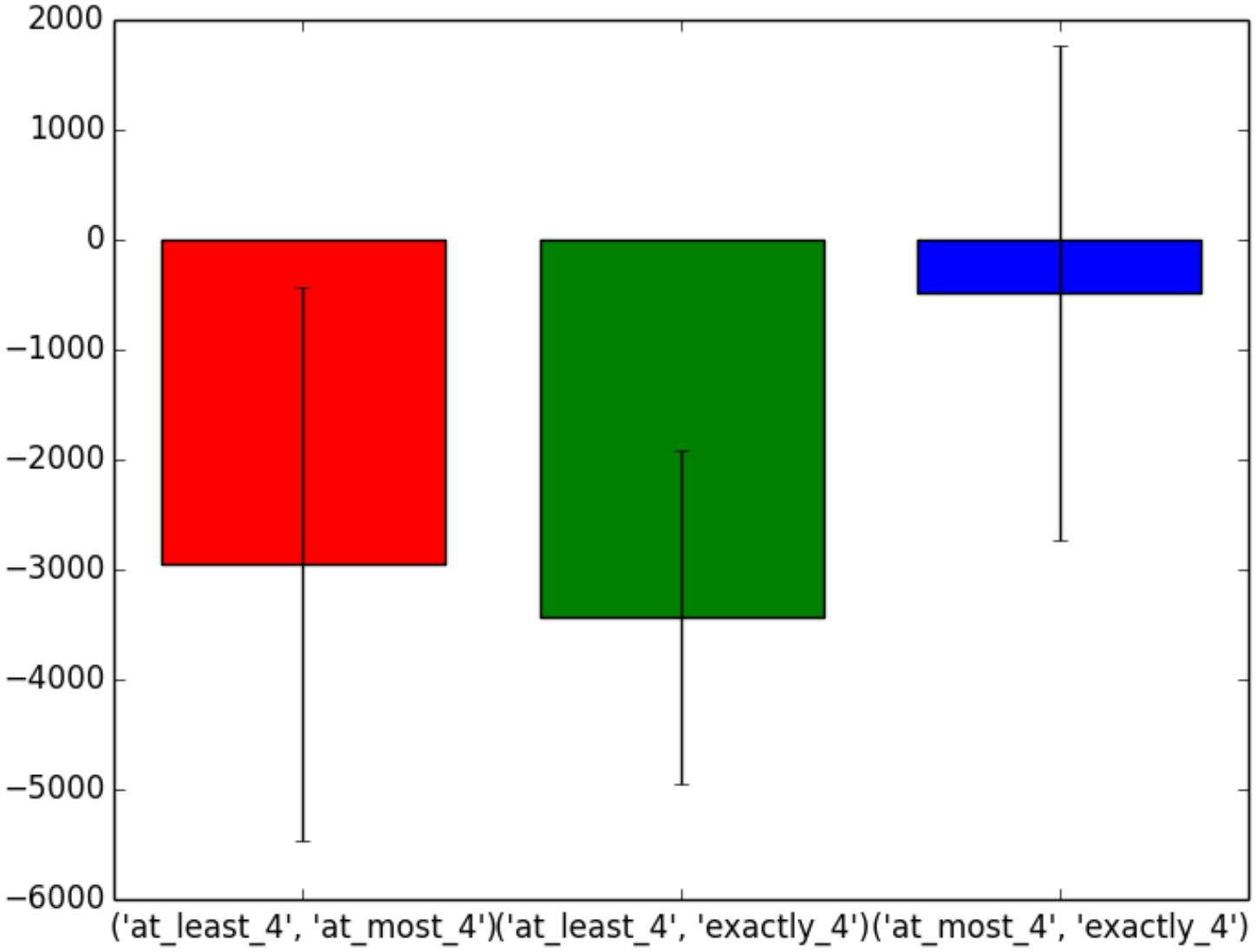
All simple determiners are monotone.

- Quantifiers: ≥ 4, ≤ 4, $= 4$

Monotonicity: Results



Monotonicity: Results



Monotonicity: Discussion

- ≥ 4 easier than $= 4$: ✓

Monotonicity: Discussion

- ≥ 4 easier than $= 4$: ✓
- ≥ 4 easier than ≤ 4, and ≤ 4 not easier than $= 4$: ???

Monotonicity: Discussion

- ≥ 4 easier than $= 4$: ✓
- ≥ 4 easier than ≤ 4, and ≤ 4 not easier than $= 4$: ???
  - Upward monotone are generally cognitively easier than downward [Just and Carpenter 1971; Geurts 2003; Geurts and Slik 2005; Deschamps et al. 2015]

Monotonicity: Discussion

- ≥ 4 easier than $= 4$: ✓
- ≥ 4 easier than ≤ 4, and ≤ 4 not easier than $= 4$: ???
  - Upward monotone are generally cognitively easier than downward [Just and Carpenter 1971; Geurts 2003; Geurts and Slik 2005; Deschamps et al. 2015]
  - $= 4$ is a conjunction of monotone quantifiers

Monotonicity: Discussion

- ≥ 4 easier than $= 4$: ✓
- ≥ 4 easier than ≤ 4, and ≤ 4 not easier than $= 4$: ???
  - Upward monotone are generally cognitively easier than downward [Just and Carpenter 1971; Geurts 2003; Geurts and Slik 2005; Deschamps et al. 2015]
  - $= 4$ is a conjunction of monotone quantifiers
- A better measure of learning rate than convergence point?

Experiment 2: Quantity

Quantity Universal

All simple determiners are quantitative.

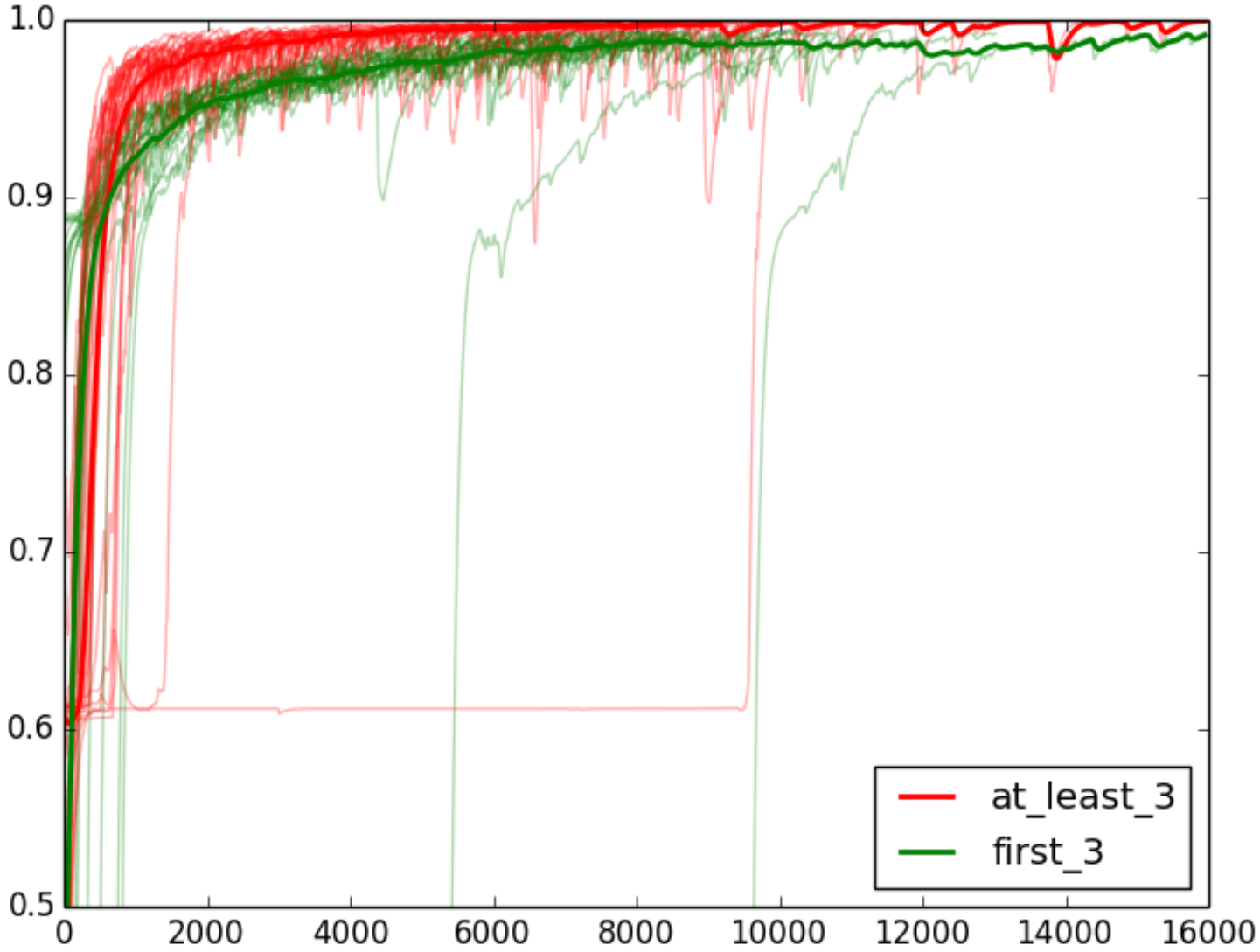
Experiment 2: Quantity

Quantity Universal

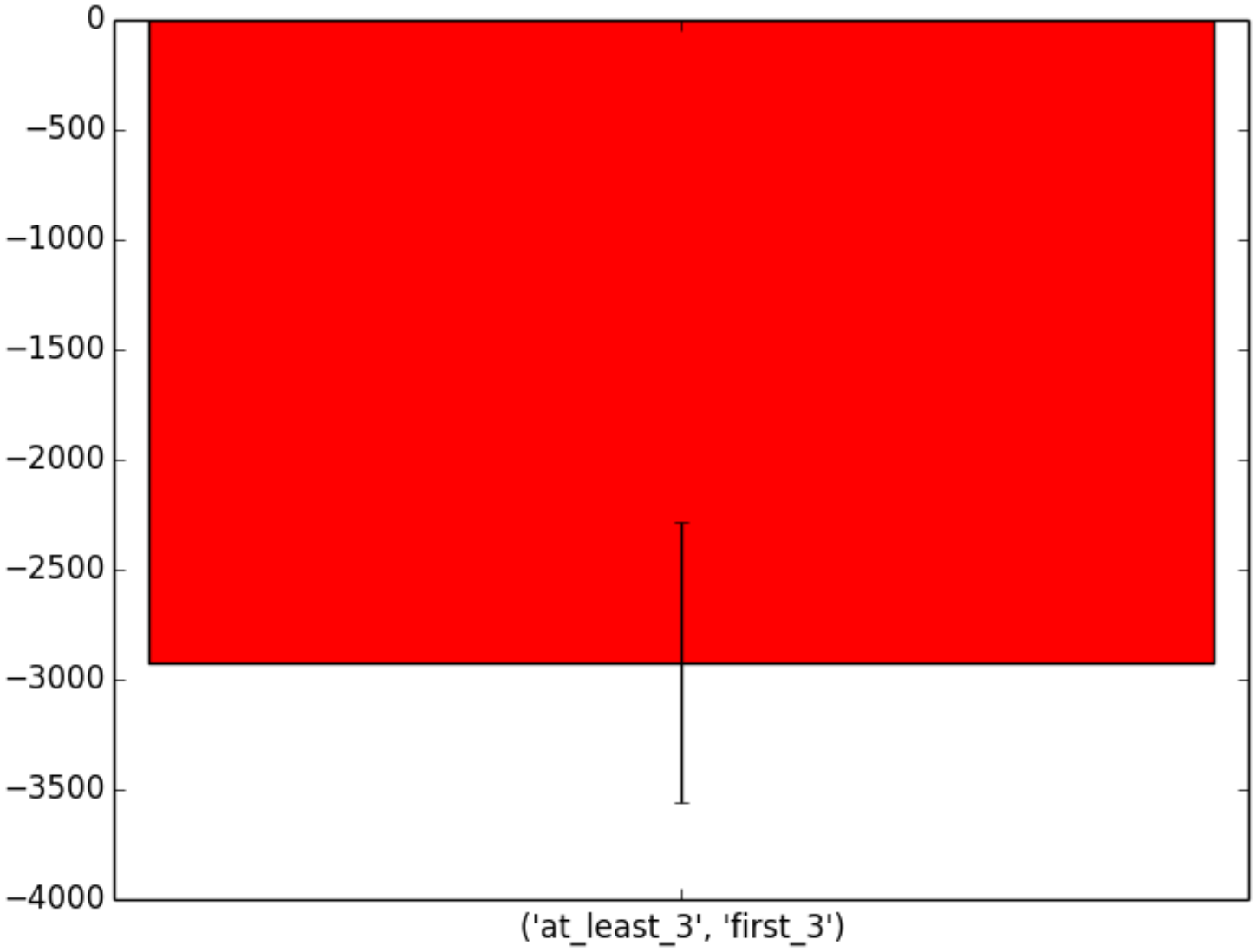
All simple determiners are quantitative.

- Quantifiers: ≥ 3, first3

Quantity: Results



Quantity: Results



Quantity: Discussion

- Very promising initial results

Quantity: Discussion

- Very promising initial results
- Harder to learn an order-sensitive quantifier than one that is only sensitive to quantity

Experiment 3: Conservativity

Conservativity Universal (Barwise and Cooper 1981)

All simple determiners are conservative.

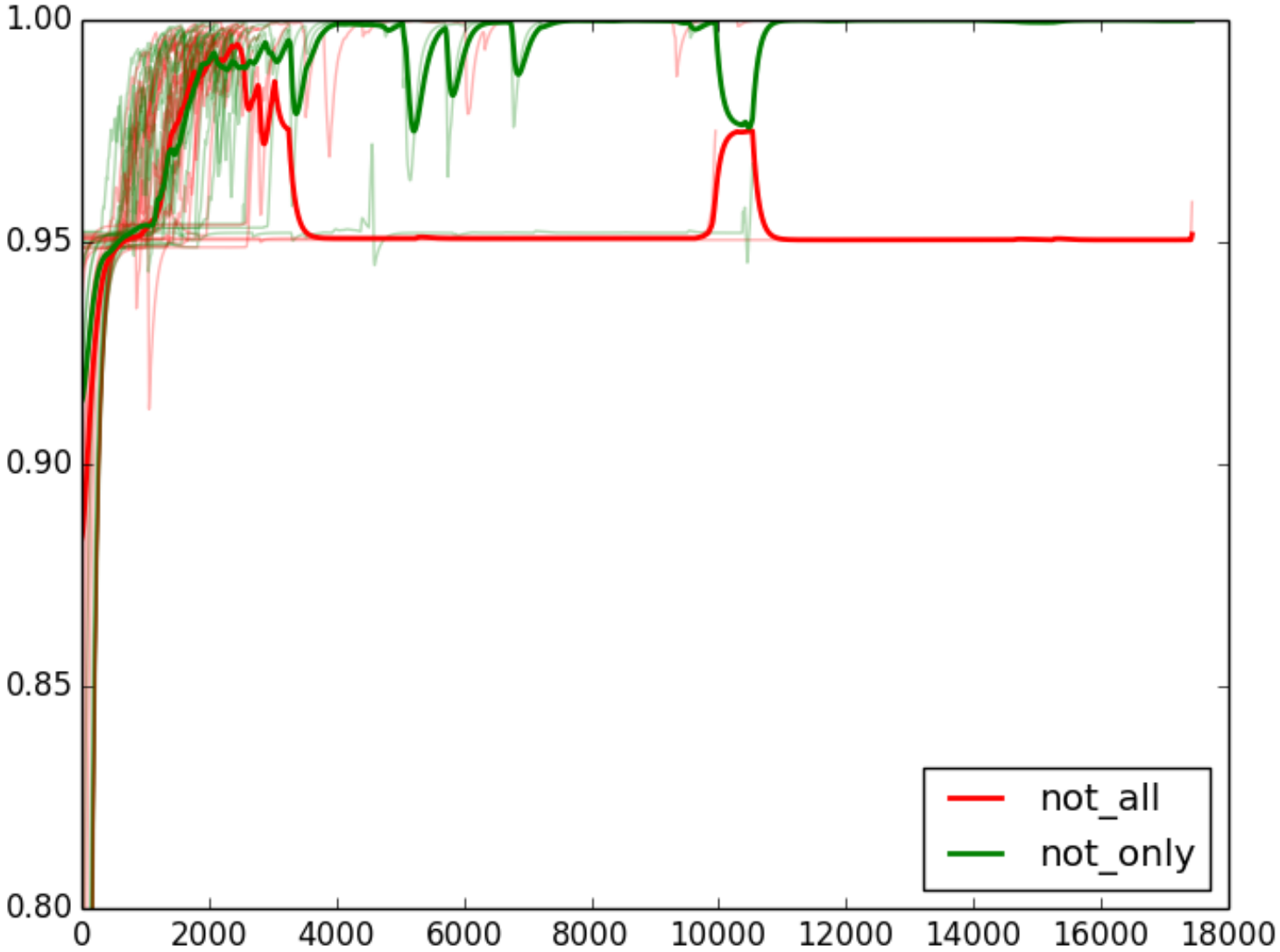
Experiment 3: Conservativity

Conservativity Universal (Barwise and Cooper 1981)

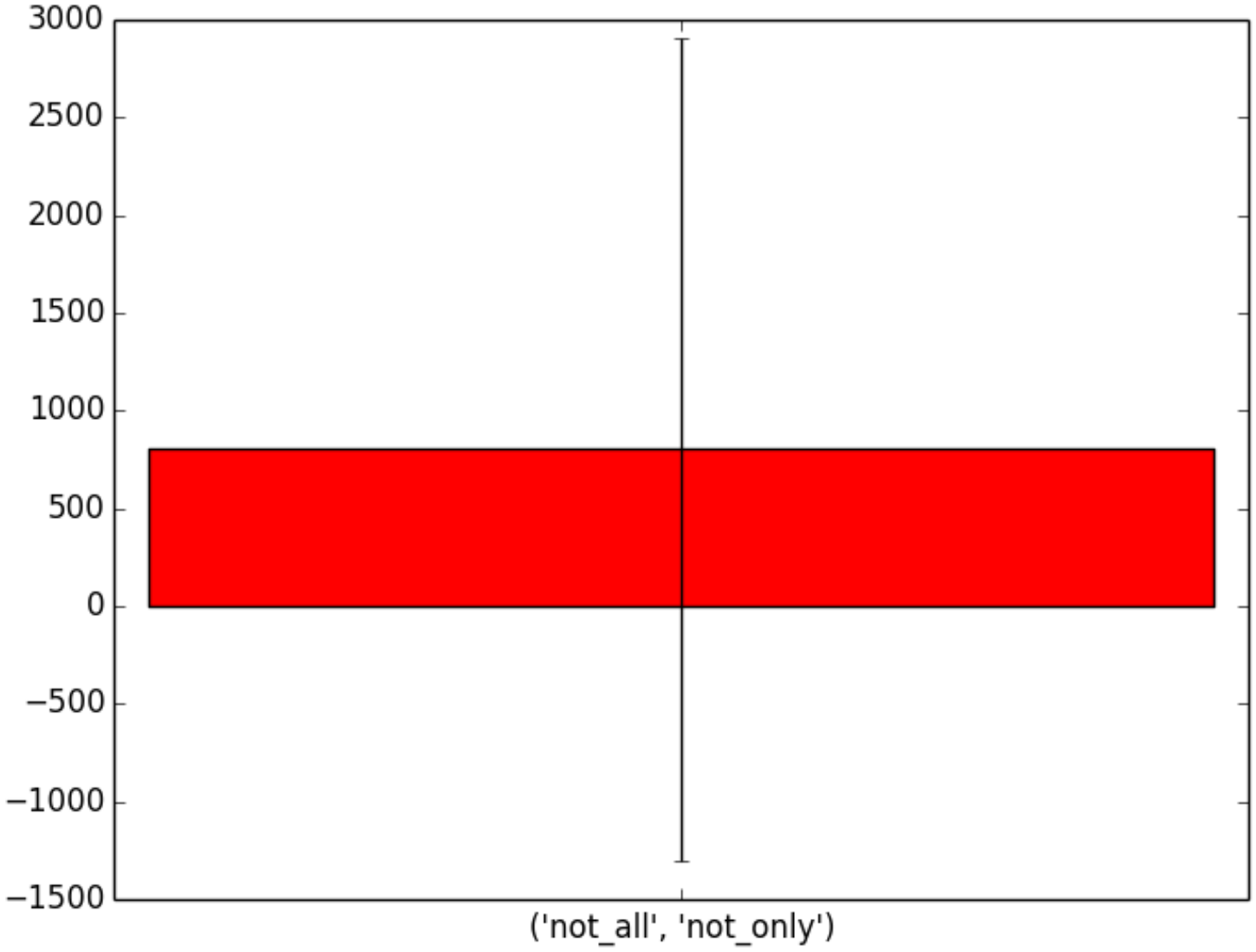
All simple determiners are conservative.

- Quantifiers: nall, nonly
  [Motivated by Hunter and Lidz 2013]

Conservativity: Results



Conservativity: Results



Conservativity: Discussion

- No way of 'breaking the symmetry' between $A \cap B$ and $B \setminus A$ in this model

Conservativity: Discussion

- No way of 'breaking the symmetry' between $A \cap B$ and $B \setminus A$ in this model
- More boldly: conservativity as a syntactic/structural constraint, not a semantic universal
  [See Fox 2002; Sportiche 2005; Romoli 2015]

Overview

Towards Meeting the Challenge

- Challenge: find a model of learnability on which quantifiers satisfying a universal are easier to learn than those that do not.

Towards Meeting the Challenge

- Challenge: find a model of learnability on which quantifiers satisfying a universal are easier to learn than those that do not.
- We showed how to train LSTM networks via backpropagation to verify quantifiers.

Towards Meeting the Challenge

- Challenge: find a model of learnability on which quantifiers satisfying a universal are easier to learn than those that do not.
- We showed how to train LSTM networks via backpropagation to verify quantifiers.
- Initial experiments show that this setup may be able to meet the challenge.

Future Work

- More and larger experiments

Future Work

- More and larger experiments
- Different models, including baselines for this task

Future Work

- More and larger experiments
- Different models, including baselines for this task
- More realistic visual input (see Sorodoc, Lazaridou, et al. 2016; Sorodoc, Pezzelle, et al. 2017)

Future Work

- More and larger experiments
- Different models, including baselines for this task
- More realistic visual input (see Sorodoc, Lazaridou, et al. 2016; Sorodoc, Pezzelle, et al. 2017)
- Tools to 'look inside' the black box and see what the network is doing

The End

Thank you!