When domain general learning fails and when it succeeds; Identifying the contribution of domain specificity

Lisa Pearl & Jeff Lidz
CNL Lunch Talk
November 30, 2006

Road Map

Learning
Learning Theories, Domain-Specificity, and Domain-Generality
Bayesian Updating

Case Studies & Models
Anaphoric One
Previous Proposals & Equal-Opportunity Bayesian Learners
Spectacular Failures & Necessary Bias

Domain-General Update Procedures

• **Probabilistic reasoning**: good for problems with noisy data or incomplete information & generally applicable to any problem space

• Key: only works over a **defined hypothesis space** (*doesn’t replace having one*)

Road Map

Learning
Learning Theories, Domain-Specificity, and Domain-Generality
Bayesian Updating

Case Studies & Models
Anaphoric One
Previous Proposals & Equal-Opportunity Bayesian Learners
Spectacular Failures & Necessary Bias
Hypothesis Spaces & Updating

- Layout of the hypothesis space and relationships between hypotheses affect how updating works.
- Updating = shifting probability between opposing hypotheses.

Look at four different hypothesis spaces:
- Non-overlapping, no initial bias
- Non-overlapping, initial bias
- Overlapping (simple), no initial bias
- Overlapping (subset-superset), no initial bias

Non-Overlapping, No Initial Bias

Two Non-Overlapping Hypotheses, Equally Probable Initially

Hypothesis A
Prob(A) = 0.5

Hypothesis B
Prob(B) = 0.5

Two Non-Overlapping Hypotheses (Equal Initial Probability), after seeing input (d, data points) that consists only of examples of A

Hypothesis A
Prob(A) = 1.0

Hypothesis B
Prob(B) = 0.0

Two Non-Overlapping Hypotheses (Equal Initial Probability), after seeing input (d, data points) that consists of 30% A examples and 70% B examples

Hypothesis A
Prob(A) = 0.3

Hypothesis B
Prob(B) = 0.7
Two Non-Overlapping Hypotheses, With Initial Bias for Hypothesis A

Hypothesis A
Prob(A) = 0.7

Hypothesis B
Prob(B) = 0.3

Two Non-Overlapping Hypotheses (Initial Bias for A), after seeing input (<d data points>) that consists only of examples of A

Hypothesis A
Prob(A) = 1.0

Hypothesis B
Prob(B) = 0.0

Two Non-Overlapping Hypotheses (Initial Bias for A), after seeing input (>d data points) that consists only of examples of B

Hypothesis A
Prob(A) = 0.0

Hypothesis B
Prob(B) = 1.0

Two Non-Overlapping Hypotheses (Initial Bias for A), after seeing input (>d, data points) that consists of 30% A examples and 70% B examples

Hypothesis A
Prob(A) = 0.3

Hypothesis B
Prob(B) = 0.7

Two Overlapping Hypotheses, Equally Probable Initially

Hypothesis A
Prob(A) = 0.5

Hypothesis B
Prob(B) = 0.5

Unambiguous data point update: same as non-overlapping case
Trajectory for Different Initial Biases

Overlapping (Subset-Superset), No Initial Bias

Hypothesis B
Prob(B) = 0.5

Hypothesis A
Prob(A) = 0.5

Two Overlapping Hypotheses in a Subset Relation, Equally Probable Initially

Overlapping (Subset-Superset), No Initial Bias

Hypothesis B
Prob(B) = 1.0

Hypothesis A
Prob(A) = 0.0

Two Overlapping Hypotheses in a Subset Relation, after seeing input (data points) that consists only of examples of B

But what if the target state is A?
There are no unambiguous data points for A!

How To Converge on the Subset

• Initially bias the hypothesis space so the subset has the majority of the probability (ex: Berwick (1985) - default/marked values)

• Use properties of the Bayesian updating procedure: indirect negative evidence

Size Principle (Indirect Negative Evidence) (Tenenbaum & Griffiths, 2001)

• Size principle: uses the layout of the hypothesis space to favor the subset hypothesis A when encountering an ambiguous data point

• Two ways to describe size principle logic:
  – Likelihood of given ambiguous data point d
  – Learner expectation of set of data points d_1, d_2, ..., d_n
Size Principle: Logic via Likelihood of $d$

- Suppose the learner encounters an ambiguous data point $d$.
- Let the number of examples covered by subset $A$ be $a$.
- Let the number of examples covered by superset $B$ be $a + b$.

- The likelihood that $d$ was produced from $A$ is $1/a$.
- The likelihood that $d$ was produced from $B$ is $1/(a+b)$.
- So, $A$ has a higher probability of having produced $d$. Thus, $A$ is favored when encountering ambiguous data.

Size Principle: Logic via Learner’s Expectation of Data Points

- If $B$ were correct, learner should encounter some unambiguous data points for $B$.

- If only subset data points are encountered, a restriction to the subset becomes more and more likely.
- The more subset data points encountered, the more the learner is biased towards $A$.

Overlapping (Subset-Superset), No Initial Bias

Two Overlapping Hypotheses in a Subset Relation, after seeing input ($> d$; data points) that consists only of examples of $A$.

Trajectory for Different Data Types

- Prob($A$) = 1.0
- Prob($B$) = 0.0
Overlapping (Subset-Superset), No Initial Bias

Hypothesis B
Prob(B) = 0.7

Hypothesis A
Prob(A) = 0.3

Two Overlapping Hypotheses in a Subset Relation, after seeing input (> d, data points) that consists of 30% A examples and 70% B examples

Road Map

Learning
Learning Theories, Domain-Specificity, and Domain-Generality
Bayesian Updating

Case Studies & Models
Anaphoric One
Previous Proposals & Equal-Opportunity Bayesian Learners
Spectacular Failures & Necessary Bias

Anaphoric One: Adult Knowledge

One refers to strings of words that can be categorized as N'

“Jack likes this red ball and Lily likes that one.”

“Jack likes this ball and Lily likes that one.”

Anaphoric One: Logical Possibility

- Alternative Hypothesis: One refers to strings categorized as N^0.

- But this is not the adult hypothesis:
  * I met the member of Congress and you met the one of the Ballroom at Maryland club.

Syntactic Hypothesis Space

Hypothesis 1 (N^0): one is anaphoric to N^0

Hypothesis 2 (N'): one is anaphoric to N'

One = N'

NP

NP

det

this

N'

N^0

ball

det

this

adj

red

N'

N^0

ball

red

purple

bottle

ball

behind

his

back

This

This
Semantic Hypothesis Space

Hypothesis 1 (N-prop): the referent of one must have the same relevant property (ex: red) as the referent of the antecedent, indicated by the modifier in the N.

Hypothesis 2 (any-prop): the referent of one can have any property and does not necessarily need to have the relevant property of the antecedent.

Linked Hypothesis Spaces

Correct hypothesis: superset in syntax, subset in semantics

Anaphoric One: Children’s Knowledge

• Scenario 1: Children think one = N
  Prediction: Antecedent of one is not phrasal, and children indifferent to properties mentioned in the modifier (any-prop).
  "Look a red bottle! Do you see another one?"

• Scenario 2: Children think one = N*
  Prediction: Antecedent of one is phrasal, and children sensitive to properties mentioned in the modifier (N-prop).
  "Look a red bottle! Do you see another one?"


"Look! A red bottle."

"Look! A red bottle."

"Do you see another one?"

- Scenario 2 wins: Children think one = N'.
  - 18-month old infants have looking preference for red bottle.
  - Sensitive to properties mentioned in modifier (‘red’).
  - Therefore, behaving as if one = N', referent of NP containing one has N'-property.

"Look a red bottle! Do you see another one?"

Anaphoric One: Estimated Available Input

- ~278,000 utterances, 4017 with anaphoric one.

<table>
<thead>
<tr>
<th>Utterance Type</th>
<th>Example</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous</td>
<td>I have a red ball, but Jack doesn’t have one. (Jack has a ball, but not a red ball.)</td>
<td>10</td>
</tr>
<tr>
<td>Type I Ambiguous</td>
<td>I have a red ball, and Jack has one, too. (Jack has a red ball at all.)</td>
<td>183</td>
</tr>
<tr>
<td>Type II Ambiguous</td>
<td>I have a ball, and Jack has one, too. (Jack has a ball with some number of properties.)</td>
<td>3805</td>
</tr>
<tr>
<td>Ungrammatical</td>
<td>you must be need one.</td>
<td>20</td>
</tr>
</tbody>
</table>

Unambiguous Data

Unambiguous data points indicate that the linguistic antecedent of one must be N’ [red ball].

Example utterance & world pairing:

“I have a red ball, but Jack doesn’t have one.”
Jack has a ball, but it does not have the property red.

- one must refer to red ball, and not to ball.

Type I Ambiguous Data

- Type I Ambiguous data points do not distinguish between one anaphoric to N’ and one anaphoric to N".
  Also, they have two choices for N ‘ - red ball or ball.

- Ex: “I have a red ball, but Jack doesn’t have one.”
  Situation: Jack has no ball at all.
  - He doesn’t have a ball, or he doesn’t have a red ball?

- Ex: “I have a red ball, and Jack has one, too.”
  Situation: Jack has a red ball.
  - He has a ball, or he has a red ball?

Type II Ambiguous Data

- Type II Ambiguous data points do not distinguish between one anaphoric to N’ and one anaphoric to N". The only string available is ball, however.

- Ex: “I have a ball, but Jack doesn’t have one.”

- Ex: “I have a ball, and Jack has one, too.”

Ungrammatical Data

- Is uninformative about what one refers to.

- Ex: “He must be needs one.”
Anaphoric One: Data Recap

- ~278,000 utterances, 4017 with anaphoric one

<table>
<thead>
<tr>
<th>Utterance Type</th>
<th>Example</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous</td>
<td>I have a red ball, but Jack doesn’t have one. (Jack has a ball, but not a red ball.)</td>
<td>10</td>
</tr>
<tr>
<td>Type I Ambiguous</td>
<td>I have a red ball, and Jack has one, too. (Jack has a red ball at all)</td>
<td>183</td>
</tr>
<tr>
<td>Type II Ambiguous</td>
<td>I have a ball, and Jack has one, too. (Jack has a ball with some number of properties.)</td>
<td>3805</td>
</tr>
<tr>
<td>Ungrammatical</td>
<td>You must be need one</td>
<td>20</td>
</tr>
</tbody>
</table>

- 10 unambiguous data points is still pitifully few…

Regier & Gahl (2004)

- Use indirect evidence: the type I ambiguous data (“I have a red ball, and Jack has one, too.”)
- Adult preference for larger N’ (red ball) will lead one to refer to a red ball every single time

Regier & Gahl (2004): Size Principle Logic

- If one always refers to a red ball (and so to red ball), learner uses size principle to converge on the subset as the correct hypothesis one = red ball; red ball ==> red ball

Size Principle Logic

- If only subset data points are encountered, a restriction to the subset becomes more and more likely.
- The more subset data points encountered, the more the learner is biased towards A.


- Size principle logic lets learner converge on correct hypothesis without recourse to implicit biases or knowledge
- Learner simply uses Bayesian updating logic
Implicit Biases Revealed

• Bias 1: Only some data used as intake (not all)
  – Unambiguous and Type I Ambiguous
  – Type II Ambiguous ignored

• Bias 2: Only semantic hypothesis space considered (and thus only semantic data points)
  – Syntactic hypothesis space ignored

Equal-Opportunity Bayesian Learner

• A Bayesian learner truly without biases: Equal-Opportunity Bayesian Learner (EO Bayesian learner)

• Uses all available data (Unambiguous, Type I, and Type II Ambiguous)

• Uses both syntactic and semantic data points (recognizing that there are two linked hypothesis spaces)

Hypothesis Spaces

• Syntax (antecedent of one): $N^0$ vs. $N'$

• Semantics (referent of one in the world): $N'$-prop vs. any-prop

Linked Hypothesis Spaces

• The problem for the size principle: no hypothesis is the subset across domains

• Correct hypothesis: superset in syntax, subset in semantics
EO Bayesian Learner: Updating

- Initial State: both hypotheses are equiprobable in both syntax & semantics
- Update probabilities within each domain, based on data type observed
- Update across domains, because hypothesis spaces are linked

EO Bayesian Learner: Syntax Unambiguous Data Update

- Unambiguous data (10 of 4017 data points)
- Intuition: 1 added to numerator since learner is fully confident that unambiguous data point signals N' hypothesis

EO Bayesian Learner: Updating Syntax

- 2 hypotheses: N' and N₀
- Track pₙₙ (p₀ₙ = 1 - pₙₙ)
- Initial state: pₙₙ = 0.5

EO Bayesian Learner: Syntax Unambiguous Data Update

- Unambiguous data (10 of 4017 data points)
- Utterance: “…red ball…one…”
- World: referent of one has property red
- Intuition: 1 added to numerator since 1 data point seen

EO Bayesian Learner: Syntax Unambiguous Data Update

- Unambiguous data (10 of 4017 data points)
- Utterance: “…red ball…one…”
- World: referent of one has property red
- Intuition: 1 added to denominator since 1 data point seen
EO Bayesian Learner: Syntax Type II Ambiguous Data Update

- Type II Ambiguous data (3805 of 4017 data points)

\[ P_N = \frac{P_N^{old} \cdot t + ????}{t + 1} \]

Intuition: number added should be less than 1, since learner is not certain that type II ambiguous data point signals \( N^0 \) hypothesis.

Utterance: “…ball…one…”
World: referent of \textit{one} \text{ may have} property red (and other properties)

EO Bayesian Learner: Partial Confidence Value

- Partial confidence value \( p_{N|a} \) is based on the fact that the utterance has only a noun as the possible antecedent.

\[ …ball…one… \]

- Noun is compatible with \( N' \) hypothesis. Means Noun-only string chosen from all possible \( N' \) strings. So, depends on likelihood of choosing a Noun-only string from all possible \( N' \) strings: \( p_n \) from \( N' \).

Example Update for Type II Ambiguous

\[ p_N = 0.5, \ t = 4017, \ p_{N|a} = 0.25 \]

\[ p_N = \frac{0.5 \cdot 4017 + 0.2}{4017 + 1} = 0.499925 \] (slight bias for \( N^0 \))

Note: majority of data is type II ambiguous (modifier-less antecedent). Every time learner sees one, learner is biased towards wrong answer. Small biases can add up over time.
EO Bayesian Learner: Syntax Type I Ambiguous Data Update

- Type I Ambiguous data (183 of 4017 data points)

\[ p_N = p_{N_{\text{old}}} \frac{t + 1}{t + 1} \]

Intuition: number added should be less than 1, since learner is not certain that type I ambiguous data signal hypothesis \( N' \).

Utterance: “…red ball…one…”

World: referent of one has property red (and other properties)

EO Bayesian Learner: Updating Semantics

- 2 hypotheses: \( N'_{\text{prop}} \) and any-prop
- Track \( p_{N'_{\text{prop}}} \), \( p_{\text{any-prop}} = 1 - p_{N'_{\text{prop}}} \)
- Initial state: \( p_{N'_{\text{prop}}} = 0.5 \)
- Data types: Same-Property, Different-Property

EO Bayesian Learner: Semantics Same-Property

- Same-property data points: referent of one has same salient property as \( N' \) antecedent referent
- \(<…\text{red ball…one…}>\) \( \rightarrow \) referent of one has property red

Unambiguous data points (10) + Type I Ambiguous data points (183) + some of the Type II Ambiguous data points (???)

EO Bayesian Learner: Updating Semantics Same-Property Type II Ambiguous

- Type II Ambiguous data point

“…ball…one…” \( \rightarrow \) referent of one has some number of properties

- number of properties learner is aware of \( = c \)
- Likelihood that referent of one coincidentally has salient property that referent of antecedent has is \( 1/c \)
**EO Bayesian Learner: Updating Semantics Same-Property**

- **Same-property** data points: referent of *one* has same salient property as N' antecedent referent
  
  (red ball...) -> referent of *one* has property red
  
  Unambiguous data points (10) + Type I
  
  Ambiguous data points (183) + some of the Type II Ambiguous data points (3805*1/c)

**EO Bayesian Learner: Updating Semantics Different-Property**

- **Different-property** data points: referent of *one* has different salient property than N' antecedent referent
  
  Value added is partial confidence value, \( p_{N-'prop|d} \), which will be < 1. Same-property data point is consistent with any-property hypothesis. Partial confidence value depends on the likelihood of choosing same-property from all properties (1/c).

**EO Bayesian Learner: Semantics Same-Property Data Update**

- **Same-Property** data
  
  (193+3805*1/c of 4017 data points)

  \[
  P_{N-'prop|d} = \frac{P_{N-'prop|d \text{ old}} t + P_{N-'prop|d \text{ new}}}{t + 1}
  \]

**EO Bayesian Learner: Semantics Different-Property Data Update**

- **Different-Property** data
  
  (3805*(c-1)/c of 4017 data points)

  \[
  P_{N-'prop|d} = \frac{P_{N-'prop|d \text{ old}} t + 0}{t + 1}
  \]

Value added to numerator is 0, since different-property data point is not compatible with N'-prop hypothesis. Learner has no confidence that this data point indicates N'-prop hypothesis.
EO Bayesian Learner: Updating Linked Domains

- Hypothesis spaces are linked
- Any data point impacting one hypothesis should also have an effect on the other
Simulating an EO Bayesian Learner

- Syntax:
  - Need value for $p_n$ from $N'$
  - Note: the higher this value, the more biased towards $N'$ the learner is for type II ambiguous data
  - We’ll be generous and define strings in $N'$ categorically, instead of by individual vocabulary items
  - $N'$ strings = {Noun, Adjective Noun, Noun PP, Adjective Noun PP}
  - Ex: “ball”, “red ball”, “ball behind his back”, “red ball behind his back”

  \[ p_n \text{ from } N' = \frac{1}{4} \]

Simulating an EO Bayesian Learner

- Semantics:
  - Need value for $c$
  - $c$ is number of categories in the world learner is aware of
  - Note: the smaller $c$ is, the more the learner is biased towards the $N'$-prop hypothesis for a same-property data point. We’ll be generous and make $c$ small.
  - Let $c = 5$ ({red, purple, nice, little, behind his back})
Road Map

Learning
- Learning Theories, Domain-Specificity, and Domain-Generality
- Bayesian Updating

Case Studies & Models
- Anaphoric One
- Previous Proposals & Equal-Opportunity Bayesian Learners
- Spectacular Failures & Necessary Bias

EO Bayesian Learner Results
- EO Bayesian learner ends up with probability of the correct grammar = .0361

EO Bayesian Learner in LWF experiment
- EO Bayesian Learner

"Look! A red bottle."

EO Bayesian Learner in LWF experiment
- EO Bayesian Learner

"Look! A red bottle."

EO Bayesian Learner in LWF experiment
- EO Bayesian Learner

"Do you see another one?"

EO Bayesian Learner in LWF experiment
- EO Bayesian Learner

(a) Antecedent is $N'$ or $N^0$? (Use $p_N$)

- $N^0$  $1-p_N$
- $N'$  $p_N$
If the antecedent is "bottle" and the learner is at chance for which bottle to look at:

1. **If not red bottle, 0.5 * (1 - p_red) - consult p_null-prop to determine if one referent must have same property as antecedent referent (N-property).**
2. **If same-property restriction, then look at red bottle (same property, N-property).**
3. **If not same-property restriction, then at chance for looking at red bottle.**
4. **To determine change of looking at the red bottle, sum the probabilities of all decisions that lead to looking at the red bottle.**
EO Bayesian Learner in LWF experiment

• Baseline probability for looking at red bottle, when given 2 bottles: .5
• Probability after learning: 0.518

That’s barely a 2% change above baseline (1/25 of baseline).

Real Learners in LWF experiment

Change from baseline: 15%, or 1/3 of baseline

EO Bayesian Learner Summary

• EO Bayesian Learner doesn’t converge on the correct grammar
• EO Bayesian Learner doesn’t behave as real learners do

Therefore, EO Bayesian Learner is not a good model of how children learn.

Less Generous Estimates for EO Bayesian Learner

• Several places where we made generous estimates of the parameters involved in the model. (Ex: p_n from N’ · c)
• This gives an overestimation of the probability an EO Bayesian learner would converge on the correct grammar.

Less Generous Estimates for EO Bayesian Learner

• p_n from N’
  – Strings defined categorically (Noun, Adjective Noun, Noun PP, etc.). Previously 1/4 (0.25).
  – Let strings be defined over vocabulary items (ball, red ball).
  – MacArthur CDI suggests 18-month olds know at least 49 adjectives and 247 Nouns, so conservative estimate of N’ strings is 49*247 (Noun Adjective combinations).
  – p_n from N’ now = .02041

Less Generous Estimates for EO Bayesian Learner

• c: number of properties learners are aware of
  – Previously 5.
  – MacArthur CDI suggests 18-month olds know at least 49 adjectives, which means they ought to know at least 49 properties.
Less Generous Results for EO Bayesian Learner

• Probability of converging on the correct syntactic and semantic hypothesis = 0.0139.

• Probability of looking at red bottle in LWF experiment = 0.507 (change from baseline of 7%, compared to real learners 15%)

Not like real learners…

The Effects of Filtering

• EO Bayesian Learner:
  – Defined hypothesis spaces, probabilistic updating
  – No filters on data intake
  – Failed badly

• Putting the Regier & Gahl filters back:
  – Use only semantic data
  – Use only unambiguous & type I ambiguous data (ignore type II ambiguous data - NPs with no modification)

The Effects of Filtering

The Effects of Filtering

Filtering Summary

• The learner does best when using both syntactic & semantic data to update, and when the learner ignores type II ambiguous data

Ignoring Type II Ambiguous Data

In order to ignore it, learner must have some way to identify type II ambiguous data

Filter to ignore type II ambiguous data should be the result of some other principled learning strategy

Proposal to Derive the Filter

• Principle: learning happens when there is uncertainty (Shannon, 1948; Gallistel, 2001; Gallistel, forthcoming)

• Suppose learner comes equipped with constraint on available representation: no anaphora to X0 categories (Hornstein & Lightfoot, 1981)

• Current problem solved: one = N

• Different problem: which N’ --> ball or red ball?
Proposal to Derive the Filter
• *Ball* and *red ball* have different consequences in the semantic domain (*any-prop* vs. *N*-prop)

• Relevant data: utterances where there is a choice between two (or more) *N*’ antecedents (*Unambiguous* and *Type I Ambiguous*) - learner has uncertainty about which *N*’ is antecedent
  “Look, a red ball! There’s another one.”

• Irrelevant data: everything else (*Type II Ambiguous*) - learner has no uncertainty about which *N*’ is antecedent
  “Here’s a ball. Give me another one, please.”

Conclusions: Learning Theory Recap
• Learning theory: not just one indivisible piece

• Three parts:
  – Definition of the hypothesis space
  – Definition of the data used as intake (filtering)
  – Procedure used to update learner’s beliefs about opposing hypotheses

In principle, any of these components could be domain-specific or domain-general

Conclusions
• Bayesian Learner fails on anaphoric *one* without filtering (& principled way to derive filter involves having constraints on the hypothesis space)

• Linked hypothesis spaces intensifies effect of learning (really good or really bad)

• Linked hypothesis spaces may mean there’s no subset hypothesis across domains, which nullifies big advantage of Bayesian updating

THE END