(First) Language Acquisition

Ted Briscoe

Computer Laboratory
Natural Language and Information Processing Group
University of Cambridge

ENS, Paris
Mar 2014

Language Learning

- Reflections on Language (1975), "To come to know a human language would be an extraordinary intellectual achievement for a creature not specifically designed to accomplish this task."
- Universal Grammar: Innate knowledge of grammar which constrains learnable grammars to a finite set specified parametrically e.g. OV/VO? $\wedge$ RelN/NRel? $\wedge$...
- Learning: is setting parameters on the basis of exposure to form:meaning pairs, but noise:
- Daddy threw you the red sock
  give′(daddy′ you′ x) $\wedge$ red′(sock′(x))
- Parameter Indeterminacy: VO-v2 or OV+v2?
  The red sock threw Daddy you

Language Learning

- Reflections on Language (1975), "To come to know a human language would be an extraordinary intellectual achievement for a creature not specifically designed to accomplish this task."
- Universal Grammar: Innate knowledge of grammar which constrains learnable grammars to a finite set specified parametrically e.g. OV/VO? $\wedge$ RelN/NRel? $\wedge$...
- Learning: is setting parameters on the basis of exposure to form:meaning pairs, but noise:
- Daddy threw you the red sock
  $give'(daddy'\ you'\ x) \wedge red'(sock'(x))$
- Parameter Indeterminacy: VO-v2 or OV+v2?
  The red sock threw Daddy you

LAD and P&P

Innate Language Acquisition Device (LAD)
– Universal Grammar (UG), Parser, Learning Procedure

Learning Procedure = Parameter setting (finite set of binary-valued independent parameters plus UG define all possible human grammars/languages)

Space of possible human languages is finite and vast:
20Ps = 1048576, 30Ps = 1.073741e+09 grammars

Reluctance to weaken learning to approximately correct

No algorithm for learnability in the limit from positive only examples

Trigger sentences = contextually determinate data for parameter setting (circumvents problems of 'evidence' / uncertainty, etc)

‘Minimalist’ Parametric Theory (Baker, Roberts et al.)

- (Int./Ext.) Merge is in UG/FLN (= CCG A,C,P...)
- Some linguistic features are in UG/FLN (= CCG Att:Val)
- Parameters relate to features (default absent/off)
- Parameters naturally define hierarchies (Decision Trees)
- Number of parameters = lg. (learning) complexity
- Parameter (re)setting = lg. change
- Macro/Micro/Meso/Nano-Parameters (head-initial/final - lexical irregularity)
- Input Generalisation / Feature Economy (none-all-some-one)
- Parameter setting is lexically-driven (observed)
- Still no theory of parameter (re)setting / learning (= LAgt LP(UG))

(Bayesian) Parametric Learning

- Bayes Rule: $P(h \mid i) = \frac{P(h)P(i|h)}{P(i|h\in\ H)}$
- Single Binary-valued Parameter: $X_0 vs. X_1$
- Input, i, 00011
- Reinforcement: $X_0 + i_0 - i_1$
- MLE/RF: $\frac{X_0}{X_0+X_1}$
- Beta Distribution + Binomial: $\frac{\alpha_0+X_0}{\alpha_0+X_0+\alpha_1+X_1}$
- Dirichlet / Pitman-Yors + Multinomial
- e.g.s $\frac{1}{2}, \frac{1}{5}, \frac{1}{50} \ldots$

No / Uniform / Informative / Accurate Prior? – Strength?
Param. Setting? = Freq. Boosting / Preserving / Averaging? – Selective

(Bayesian) Parametric Learning

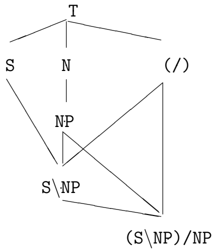
- Bayes Rule: $P(h \mid i) = \frac{P(h)P(i|h)}{P(i|h \in H)}$
- Single Binary-valued Parameter: $X_0 vs. X_1$
- Input, i, 00011
- Reinforcement: $X_0 + i_0 - i_1$
- MLE/RF: $\frac{X_0}{X_0 + X_1}$
- Beta Distribution + Binomial: $\frac{\alpha_0 + X_0}{\alpha_0 + X_0 + \alpha_1 + X_1}$
- Dirichlet / Pitman-Yors + Multinomial
- e.g.s $\frac{1}{2}, \frac{1}{5}, \frac{1}{50} \ldots$

No / Uniform / Informative / Accurate Prior? – Strength?
Param. Setting? = Freq. Boosting / Preserving / Averaging? – Selective

Bayesian Incremental Parameter Setting (BIPS)

- Input – finite noisy randomly-ordered form-meaning pairs (fm_n):
  Daddy gave you the sock throw′(daddy′ you′ x) ∧ sock′(x)
- Hypothesis Space – F/B A+C, L/D P + Cat. + Lex.
- Learning Bias / Occam's Razor – prior distribution on set of finite-valued parameters (A,C,P + Cat. Set):
  $p(g \in G) = \prod_{param_i \in g} p(param_i = x)$
- Lexical Parameters $p(Cat, Lexeme)$
- Incremental Learning, posterior distribution given input:
  for $0 < i < n, argmax_{g \in G}\ p(g)\ p(fm_i \mid g)$
  $p(fm_i \mid g) = \prod_{param_j \in fm_i} p(param_j)$
  $p(param_j = x) = \frac{f(param_j = x) + \alpha}{f(param_j = X) + N\alpha}$
- Parameter is (re)set if $argmax(p(param_j = x))$ (selective)

Parametric Specification of Category Sets



Finite Feature / Category Set:

NP	=	[CAT=N, BAR=1, CASE=X, PERNUM=Y]
S	=	[CAT=V, BAR=0, PERNUM=X]
\NP	=	[DIR = left, CAT=N,...]
		$S_{pernum=x} \backslash NP_{pernum=x}$
		$S \backslash NP_{pernum=3sg} \sqcap NP_{case=nom} = NP_{3sg,nom}$

Parameters in Type-driven HPSG / Construction Grammar

- (Default) Inheritance via (default) unification
- A grammar is a set of Constraints (CON)
- CON contains (Sub)Type Inheritance & Path Value Specifications (PVSs)
- $Verb \sqsupset\ IntransVb \sqsupset\ TransVb$
- $TransVb\ ARG2 =_d NP$
- $Rain \sqsubset\ IntransVb\ ARG1 =_d NP_{IT}$
- Parameters = non-UG part of CON associated with Probabilities/Settings

The Locally Maximal Grammar (none-all-some-one)

$\forall pPVS_i \in \; CON(Supertype_j, \sqsubset)$
$\forall pPVS_k \in \; Subtypes_l \; of \; Supertype_j$
if

$$\mid pPVS_k = 1 \in \; Subtypes_l \mid \; > \; \mid pPVS_k = 0 \in \; Subtypes_l \mid$$

then

$$P(pPVS_i = 1) = \frac{\sum P(pPVS_k{=}1) \in \; Subtypes_l}{|pPVS_k{=}1 \in \; Subtypes_l|}$$

(and vice-versa)
else
if

$$\frac{\sum P(pPVS_k{=}1) \in \; Subtypes_l}{|PVS_k{=}1 \in \; Subtypes_l|} > \frac{\sum P(pPVS_k{=}0) \in \; Subtypes_l}{|pPVS_k{=}0 \in \; Subtypes_l|}$$

then

$$P(pPVS_i = 1) = \frac{\sum P(pPVS_k{=}1) \in \; Subtypes_l}{|PVS_k{=}1 \in \; Subtypes_l|}$$

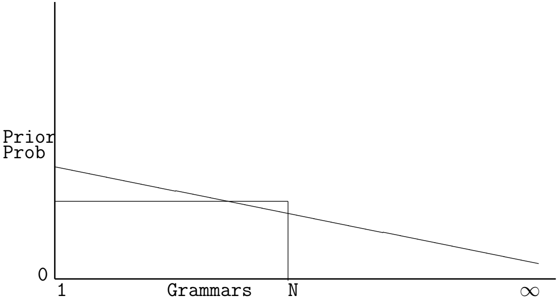
(and vice-versa)

Chomskyan vs. Bayesian Learning

Learning Universal: Irregularity correlated with frequency

go+ed / went, ((S\IT)/NP)/S annoy, bother,...

Convergent Evolution: lrng biases walk thru' parameter space



A Language

Lexicon:

Kim : NP : kim′

Sandy: NP : sandy′

Paris: NP : paris′

kissed : $(S\backslash NP)/NP$: λ y,x kiss′(x y)

in : $((S\backslash NP)\backslash(S\backslash NP))/NP$: λ y,P,x in ′(y P(x))

. . .

Grammar:

Forward Application (FA):

$X/Y\ Y \Rightarrow X \quad \lambda\, y\, [X(y)]\, (y) \Rightarrow X(y)$

Backward Application (BA):

$Y\ X\backslash Y \Rightarrow X \quad \lambda\, y\, [X(y)]\, (y) \Rightarrow X(y\ )$

A Language

Lexicon:

Kim : NP : kim′

Sandy: NP : sandy′

Paris: NP : paris′

kissed : (S\NP)/NP : λ y,x kiss′(x y)

in : ((S\NP)\(S\NP))/NP : λ y,P,x in ′(y P(x))

. . .

Grammar:

Forward Application (FA):

X/Y Y $\Rightarrow$ X λ y [X(y)] (y) $\Rightarrow$ X(y)

Backward Application (BA):

Y X\Y $\Rightarrow$ X λ y [X(y)] (y) $\Rightarrow$ X(y)

A Derivation

```
Kim    kissed             Sandy      in                          Paris
NP     (S\NP)/NP          NP         ((S\NP)\( S\ NP))/NP        NP
kim′   λ y,x kiss′(x y)   sandy′     λ y,P,x in′(y P (x))        paris′
       ------------------ FA          ----------------------- FA
       S\NP                          (S\NP)\(S\ NP)
       λ x kiss′(x sandy′)           λ P,x in′(paris′ P(x))
       ----------------------------------------- BA
       S\NP
       λ x in′(paris′ kiss′(x sandy′))
--------------------------- BA
S
in′(paris′ kiss′(kim′ sandy′))
```

. . . in Paris on Friday by the Eiffel Tower . . .

Another Language

Lexicon:

Ayse : NP : kim′

Fatma'yi: NP_{acc} : sandy′

Paris: NP : paris′

gordu : $(S\backslash NP)\backslash NP_{acc}$: λ y,x see′(x y)

+de : $((S\backslash NP)/(S\backslash NP))\backslash NP$: λ y,P,x in′(y P(x))

...

Grammar:

Composition (C):

$X/Y \; Y/Z \Rightarrow X/Z \quad \lambda\, y\, [X(y)]\; \lambda\, z\, [Y(z)] \Rightarrow \; \lambda\, z\, [X(Y(z))]$

Another Language

Lexicon:

Ayse : NP : kim′

Fatma'yi: NP_{acc} : sandy′

Paris: NP : paris′

gordu : $(S\backslash NP)\backslash NP_{acc}$: λ y,x see′(x y)

+de : $((S\backslash NP)/(S\backslash NP))\backslash NP$: λ y,P,x in′(y P(x))

...

Grammar:

Composition (C):
$X/Y \; Y/Z \Rightarrow X/Z \quad \lambda y\,[X(y)]\; \lambda z\,[Y(z)] \Rightarrow \lambda z\,[X(Y(z))]$

Another Derivation

Ayse	Fatma'yi	Paris	+de	gordu
NP	NP_{acc}	NP	$((S\backslash NP)/(S\backslash NP))\backslash NP$	$(S\backslash NP)\backslash NP_{acc}$
kim′	sandy′	paris′	λ y,P,x in′(y P(x))	λ y,x see′(x y)

-------------------- BA (Paris +de)

$(S\backslash NP)/(S\backslash NP)$

λ P,x in′(paris′ P(x))

---------------------------------- C (with gordu)

$(S\backslash NP)\backslash NP_{acc}$

λ y,x in′(paris′ see′(x y))

---------------- BA (with Fatma'yi)

$S\backslash NP$

λ x in′(paris′ see′(x sandy′))

------------- BA (with Ayse)

S

in′(paris′ see′(kim′ sandy′))

An Unlikely Language

Lexicon:
Kim : NP : kim′
Sandy: NP : sandy′
Paris: NP : paris′
kissed : (S\NP)/NP : λ y,x kiss′(x y)
in : ((S\NP)\(S\NP))/NP : λ y,P,x in ′(P(x) y)
see : S\NP)\NP : λ y,x see′(x y)
+on : ((S\ NP)/(S\ NP))\NP : λ y,P,x on′(P(x) y)
. . .
An Unlikely Derivation
Kim Friday +on Sandy see in Paris
on′(friday, in′(paris′ see′(kim′ sandy′)))

An Unlikely Language

Lexicon:
Kim : NP : kim′
Sandy: NP : sandy′
Paris: NP : paris′
kissed : (S\NP)/NP : λ y,x kiss′(x y)
in : ((S\NP)\(S\NP))/NP : λ y,P,x in ′(P(x) y)
see : S\NP)\NP : λ y,x see′(x y)
+on : ((S\ NP)/(S\ NP))\NP : λ y,P,x on′(P(x) y)
. . .

An Unlikely Derivation
Kim Friday +on Sandy see in Paris
on′(friday, in′(paris′ see′(kim′ sandy′)))

The Basic Stochastic ILM

For i =1 to N,
$LAgt_1 :< lg^t, Generate(lg^t, m_i), Age(1) >$
$LAgt_2 :< lg^{t+i} = LP(UG, fm_i), Parse(lg^{t+i}, m_i), Age(0) >$

$\longrightarrow$

for i =1 to N,
$LAgt_2 :< lg^{t+N}, Generate(lg^{t+N}, m_i), Age(1) >$
$LAgt_3 :< lg^{t+N+1} = LP(UG, fm_i), Parse(lg^{t+N+1}, m_i), Age(0) >$

$\longrightarrow$

...

If N is large enough to guarantee (random) generation of a fair sample of lg^t and UG provides an uninformative or accurate prior, then $lg^t = lg^{t+N}$ for all $t+N$
If there is noise, variation or bottleneck, then prior will dominate over time (Griffiths, Kirby)

The Basic Stochastic ILM

For i =1 to N,
$LAgt_1 :< lg^t, Generate(lg^t, m_i), Age(1) >$
$LAgt_2 :< lg^{t+i} = LP(UG, fm_i), Parse(lg^{t+i}, m_i), Age(0) >$

$\longrightarrow$

for i =1 to N,
$LAgt_2 :< lg^{t+N}, Generate(lg^{t+N}, m_i), Age(1) >$
$LAgt_3 :< lg^{t+N+1} = LP(UG, fm_i), Parse(lg^{t+N+1}, m_i), Age(0) >$

$\longrightarrow$

...

If N is large enough to guarantee (random) generation of a fair sample of lg^t and UG provides an uninformative or accurate prior, then $lg^t = lg^{t+N}$ for all $t+N$
If there is noise, variation or bottleneck, then prior will dominate over time (Griffiths, Kirby)

LP Effectiveness

Number of (randomly-generated) inputs (in increments of 10) required to ensure convergence to g_t ($P = 0.99$):

Learner	**Language**							
	SVO	SVOv1	VOS	VSO	SOV	SOVv2	OVS	OSV
Uniform	60	80	70	80	70	70	70	70
SVO-Prior	60	60	60	60	60	60	80	70

Emergent Compositionality (Kirby)

- Suppose *Generate* invents f_k for m_k when not in lg^t?
- Input:

li+co+ba+gu	li+bo+ri
S	S
see′(kim′ sandy′)	kiss′(kim′ fido′)

- Acquired Lexicon:

co+ba+gu	:	S\NP	:	λ x see′(x sandy′)
bo+ri	:	S\NP	:	λ x kiss′(x fido′)
li	:	NP	:	kim′

Emergent Compositionality (Kirby)

- Suppose *Generate* invents f_k for m_k when not in lg^t?
- Input:

li+co+ba+gu	li+bo+ri
S	S
see′(kim′ sandy′)	kiss′(kim′ fido′)

- Acquired Lexicon:

co+ba+gu	:	S\NP	:	λ x see′(x sandy′)
bo+ri	:	S\NP	:	λ x kiss′(x fido′)
li	:	NP	:	kim′

Emergent Compositionality (Kirby)

- Suppose *Generate* invents f_k for m_k when not in lg^t?
- Input:

li+co+ba+gu	li+bo+ri
S	S
see′(kim′ sandy′)	kiss′(kim′ fido′)

- Acquired Lexicon:

co+ba+gu	: S\NP	: λ x see′(x sandy′)
bo+ri	: S\NP	: λ x kiss′(x fido′)
li	: NP	: kim′

Learning Procedure Desiderata

1. Realistic Input noisy, non-homogeneous input
2. Accurate parameters are set based on input
3. Selective parameters are set to most probable value
4. Generalisation resulting grammars are productive
5. Regularisation inductive bias for regularity
6. Occam's Razor grammars are minimal wrt input

Summary

- I-lgs / CCGs can be incrementally learnt via BIPS LP(UG)
- ILM over BIPS LP is stable and accurate – prior?
- Productivity (recursion) is a property of Application and Composition
- With invention in *Generate* compositionality is emergent
- Where does E-lg variation and change come from?

Reading

Baker, M *The Atoms of Language*, OUP, 2001
Biberauer, T., Holmerg, A., Roberts, I., Sheehan, M., "Complxity in Comparative Syntax: The View from Modern Parametric Theory" ms.,
www.mml.cam.ac.uk/dtal/research/recos/
Kirby, S. "Learning Bottlenecks and the Evolution of Recursive Syntax"
and
Briscoe, E.J. "Grammatical Acquisition and Linguistic Selection", in *Linguistic evolution through language acquisition: formal and computational models*, (ed.) Briscoe, E.J., Cambridge University Press, pp255-300, 2002 www.cl.cam.ac.uk/users/ejb/