Simplicity in grammar and the Bayesian evaluation measure

Ewan Dunbar

October 6, 2012
(MA)NECPhon 6: University of Maryland, College Park

Summary

(1) *Evaluation measure:* a static description of the relative "value" of grammars to the learner (as opposed to a description of *how* learning takes place)

(2) *One classic evaluation measure:* disprefer grammars that take more symbols to write down; attributes to the learner an intuition (one of several conflicting intuitions) linguists appeal to when they need to choose between different analyses

(3) *Bayesian:* Bayesian statistics uses the posterior distribution of the grammar's free parameters as a static description of relative value—so, an evaluation measure

(4) *Bayesian Occam's Razor:* An automatic preference for simpler models (Jaynes 2003, MacKay 2003); present in any hierarchical Bayesian analysis with certain properties, but usually discussed only in passing

(5) *Goal of this handout:* Do enough exegesis of the BOR effect that we feel like we understand it and can derive a symbol-counting type evaluation measure from it

(6) *Goal of this talk:* Provide a recipe for getting a BOR effect in inference that you can try to hook up to your favorite language acquisition problem

Evaluation measures

(7) Functions mapping from grammars to some set of "values" supporting a comparison of those grammars; more "valuable" grammars will be preferred by the learner, all things being equal (Chomsky 1964, Chomsky 1965, Chomsky and Halle 1965, Chomsky and Halle 1968)

(8) *Example:* "The 'value' of a sequence of rules is the reciprocal of the number of symbols in the minimal schema that expands to this sequence." (Chomsky and Halle 1968, 334)

(9) 1/ the number of symbols in the grammar is theorized to be your guide to choosing between different grammars that are all equally consistent with the data; more generally, it is supposed to trade off against consistency with the data (since, as CH recognize, the fit will never be perfect)

(10) If this were a probability distribution (it's not) it would be a prior distribution

(11) They take it to be an accident that their evaluation measure measures value in a way that corresponds to an intuitive simplicity (in terms of number theoretical elements used)

(12) In fact, to make this point frustratingly clear, they take the step of using the term "simplicity" as a technical term, just to mean "value," as in, "output of the evaluation measure"

> (13) "It is not of the slightest importance to us that the simplest grammar, in our sense, may be difficult for some linguists to read, or that it may be wasteful of printer's ink." (Chomsky and Halle 1965, 110)

(14) *We show:* given a theory and some general guidelines about how you hook it up to inference, in fact, there is a domain-general (Bayesian) reason that you would have an evaluation measure like this

Simplicity in linguistics

(15) Choice of more or less complex grammatical description for a language

(16) *Example:* deciding between $V \rightarrow \phi$ /VC—C+V and $V \rightarrow \phi$ /VC—CV

(17) This has empirical consequences, but it becomes especially relevant when it does not have consequences which are *available to the learner* (or, less interestingly, the analyst)

(18) In this case simplicity leads to a larger string-language, and thus is in this case the polar opposite of *restrictiveness*

(19) Both intuitions are appealed to by linguists (and sometimes unlucky undergraduates get one piece of advice one day and the other the next)

(20) The automatic Bayesian magic which gives restrictiveness has been discussed elsewhere (Tenenbaum 1999, Jarosz 2009)

(21) *Turns out:* The same basic facts about probability—operating in a different part of the inference, the prior, rather than the likelihood—give rise to BOR, thus, a countervailing force to the restrictiveness pressure

Model evaluation in statistics

(22) *Parameter estimation:* basic operation in statistical inference; a procedure which takes a collection of data as input, and returns a useful model intended to be good at predicting the same data.

(23) *Model evaluation:* a separate inference from parameter estimation, about higher-level "model framework" the parameters appear to fit into

(24) *Example:* Regression—predict *response* variable from some number of *predictor* variables using a linear function

(25) $$y = \theta_1 x_1 + \theta_2 x_2$$

(26) *Parameter estimation:* find a good value for $\langle \theta_1, \theta_2 \rangle$

(27) *Model evaluation:* decide whether $y = \theta_1 x_1 + \theta_2 x_2$ or $y = \theta_1 x_1$ is the right model framework

(28) *Practical motivations:* avoiding overfitting, assessing whether an effect is large and/or reliable enough to be considered "real"

(29) *Crucial:* model evaluation represents a **separate question** from parameter estimation, **even if** there appear to be two ways to approach the same question (e.g., ask whether parameter estimation recommends a value with $\theta_2 = 0$, or ask the model evaluation whether $\theta_2 \neq 0$ should be allowed)

(30) *Terminology:* we will call the goal of inference in parameter estimation *parameter values* or *model instantiations;* we will call the goal of inference in model evaluation *model frameworks*

Bayesian inference and model comparison

(31) *Bayesian inference:* Use probabilities of parameter values (given the data) to assess "value"

(32) *Bayesian model comparison:* Add an extra parameter ω to code for different model frameworks

(33) *Bayes factor:* Compare two different model frameworks by taking their posterior ratio

(34) $$\frac{\Pr[X|\omega{=}1]\,\Pr[\omega{=}1]}{\Pr[X|\omega{=}2]\,\Pr[\omega{=}2]}$$

(35) From now on, we will ignore the prior bias for one model or the other (the second ratio)

(36) *Example:* toy regression, $\Theta = \{\langle 0,0\rangle, \langle 1,0\rangle, \langle 0,1\rangle, \langle 1,1\rangle\}$

(37) We use the law of total probability and the chain rule to expand the Bayes factor

(38) We write subscripted λ for the likelihood ($\Pr\left[X|\overline{\theta}\right]$), a subscripted $p^{(1)}$ for $\Pr\left[\overline{\theta}|\omega = 1\right]$, a subscripted $p^{(2)}$ for $\Pr\left[\overline{\theta}|\omega = 2\right]$.

(39) $$\frac{\Pr\left[X|\omega = 1\right]}{\Pr\left[X|\omega = 2\right]} = \frac{\lambda_{0,0}p^{(1)}_{0,0} + \lambda_{1,0}p^{(1)}_{1,0} + \lambda_{0,1}p^{(1)}_{0,1} + \lambda_{1,1}p^{(1)}_{1,1}}{\lambda_{0,0}p^{(2)}_{0,0} + \lambda_{1,0}p^{(2)}_{1,0} + \lambda_{0,1}p^{(2)}_{0,1} + \lambda_{1,1}p^{(2)}_{1,1}}$$

(40) *Reduced model,* $\omega = 1$: $\Pr\left[\overline{\theta}|\omega = 1\right]$ is a distribution over the parameter space for which all events where $\theta_2 \neq 0$ have probability 0

(41) *Full model,* $\omega = 2$: $\Pr\left[\overline{\theta}|\omega = 2\right]$ is some distribution over the full parameter space Θ

(42) $$\frac{\Pr[X|\omega=1]}{\Pr[X|\omega=2]} = \frac{\lambda_{0,0}p_{0,0}^{(1)} + \lambda_{1,0}p_{1,0}^{(1)}}{\lambda_{0,0}p_{0,0}^{(2)} + \lambda_{1,0}p_{1,0}^{(2)} + \lambda_{0,1}p_{0,1}^{(2)} + \lambda_{1,1}p_{1,1}^{(2)}}$$

(43) *Idea:* $p_{0,0}^{(1)} + p_{1,0}^{(1)} = 1$; but $p_{0,0}^{(2)} + p_{1,0}^{(2)} \leq 1$; so the probabilities on the top will weight the *same* likelihood values (the fit of the two reduced-model parameter values) by *larger* numbers

(44) *Turns out:* Under some simple sanity assumptions about how the two model frameworks assign probability, we can see the numerator prior probabilities as scaled (scaled-*up*) versions of the same two probabilities in the denominator

(45) $$\frac{\Pr[\omega=1|X]}{\Pr[\omega=2|X]} = \frac{1}{p_{0,0}^{(2)} + p_{1,0}^{(2)}} \cdot \frac{\lambda_{0,0}p_{0,0}^{(2)} + \lambda_{1,0}p_{1,0}^{(2)}}{\lambda_{0,0}p_{0,0}^{(2)} + \lambda_{1,0}p_{1,0}^{(2)} + \lambda_{0,1}p_{0,1}^{(2)} + \lambda_{1,1}p_{1,1}^{(2)}}$$

(46) *Bayesian Occam's Razor:* Prefer the simpler model (all things being equal) just because the probabilities need to be distributed over more possible values in the complex model, and thus need to be scaled down there

(47) We go through how and when this will happen in the Appendix

(48) Ultimately, this is always a consequence of *monotonicity*—if $A \subseteq B$, then $\Pr[A] \leq \Pr[B]$, for any probability distribution; since $\Theta_1 \subseteq \Theta_2$, in this case, $\Pr[\Theta_1|\omega_2] \leq 1$—and the fact that probability distributions are *unit measures* (sum to one)

Summary of Appendix

(49) *Formula:*

- (50) Look at an inference problem.
- (51) Formulate it so that it contains a "hyperparameter" adjusting the complexity of the model
- (52) *Align:* For any two different levels of complexity, ensure you can always put some of the parameter values in correspondence across the two, at least for some subset of each
  - (53) *Conditional equality:* The correspondence needs to preserve relative probability of parameter values within the matching subsets
  - (54) *Likelihood matching:* The correspondence needs to preserve the likelihood on the current data set
- (55) *Conclude:* Prefer "simpler" values of this hyperparameter—meaning ones where there is less other stuff outside the aligned subset—all things being equal

Grammatical inference

(56) *Formula:*

- (57) Look at a grammatical inference problem.

(58) Find some atom of the theory, or some inference that would be necessary to formulate a grammar, which effectively adjusts the complexity of the grammar

(59) *Align:* For any two different levels of complexity, ensure some subset of the grammars can be put in correspondence (for example, if one level of complexity can be seen as a refinement)

  (60) *Conditional equality:* The correspondence needs to preserve relative preference for grammars within the matching subsets (i.e., adding the refinement should not change relative preferences)

  (61) *Likelihood matching:* The correspondence needs to preserve the likelihood on the current data set (the data should be insensitive to the refinement)

(62) *Conclude:* Prefer "simpler" grammars, all things being equal

Deriving a symbol-counting evaluation measure

(63) *The "value" of a sequence of rules is the reciprocal of the number of symbols in its minimal representation.* (334)

(64) *Example:* (with apologies to Chomsky and Halle 1965)

	$\mathrm{i} \rightarrow \mathrm{y}\ /\text{—}\mathrm{p}$	$\succ$	$\mathrm{i} \rightarrow \mathrm{y}\ /\mathrm{a}\text{—}\mathrm{p}$
symbols:	"6"		"7"
value:	$\frac{1}{\text{"6"}}$	$>$	$\frac{1}{\text{"7"}}$

(65) There is no sense trying to derive this evaluation measure, because it cannot be interpreted as a probability; it is not a unit (nor even finite) measure.[1]

(66) However, we can weaken (63) to something following the same intuition:

(67) *The value of a sequence of rules decreases in the number of symbols in its minimal representation.*

(68) Set up a bijection f (say, for 6-symbol and 7-symbol grammars)

(69) *Align grammars*: $f_{S,I}(G_1) = G_2$ iff G_2 is the result of inserting the symbols $S = \{s_1, \ldots, s_k\}$ at positions $I = \{i_1, \ldots, i_k\}$ in G_1.

  (70) *Conditional prior:* Assume there is only a weak substantive bias such that there should be nothing about assuming a particular refinement that changes the relative value of a rule, or a grammar.

  (71) *Likelihood:* Both the grammar and its refinement should be equally compatible with the available data.

(72) The mapped subset is the whole set of possible grammars under G_1; it is a proper subset under G_2, since there are many different bijections we could construct; thus G_1 is preferred.

(73) *Other approaches:* Need not assume that there is an inference for the total number of symbols in the grammar, but as this number remains unbounded and the grammar needs to be finite, inference will always need to do the work of limiting **something** (length of rule, length of left/right context, ...); so long as the prior cuts the grammar at its joints, the BOR will then kick in

[1] Let N_p be the number of grammars of length p (some integer $\leq |V|^p$). Then $\sum_G \Pr[G] = \sum_{p=1}^{\infty} \frac{N_p}{p} = \infty$.

Summary

(74) *Bayesian Occam's Razor:* An automatic preference for simpler models; showed (in the Appendix) when exactly this will hold; simpler means basically that there are fewer possible model instances

(75) *Point:* You are adjusting the size of the set of possible model instances when you adjust hypothesized grammars to be simpler or more complex

(76) *Recipe:* If the learner is inferring certain things in a certain way (nothing crazy), then hypothesizing that they are a Bayesian inference machine has simpler grammars as a consequence

(77) Have fun!

References

CHOMSKY, NOAM. 1964. Current Issues in Linguistic Theory. In *The Structure of Language: Readings in the Philosophy of Language*, ed. by Jerry Fodor and Jerrold Katz. Englewood Cliffs, NJ: Prentice Hall.

——. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

——, and MORRIS HALLE. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1.97–214.

——, and MORRIS HALLE. 1968. *The Sound Pattern of English*. New York, NY: Harper and Row.

JAROSZ, GAJA. 2009. Restrictiveness and Phonological Grammar and Lexicon Learning. In *43rd Annual Meeting of the Chicago Linguistic Society*.

JAYNES, E. T. 2003. *Probability Theory: The Logic Of Science*. Cambridge: Cambridge University Press.

MACKAY, DAVID. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

TENENBAUM, JOSHUA, 1999. *A Bayesian Framework for Concept Learning*. Cambridge, MA: MIT dissertation.

Appendix: Details

(78) **Model equivalence principle (MEP):** For any pair of model framework parameter values ω_1, ω_2, the prior distributions over Θ to which they correspond must be somewhere *similar* (with respect to the likelihood function).

(79) **Similarity (easy version):** Two probability distributions $\Pr_1$ and $\Pr_2$ over Θ are *similar* for some $A \subseteq \Theta$ if, for all $S \subseteq A$, $\Pr_1[S|A] = \Pr_2[S|A]$.

(80) **Similarity (harder version):** Two prior distributions $\Pr_1$ and $\Pr_2$ over Θ_1, Θ_2 are *similar* for some $A_1 \subseteq \Theta_1$, $A_2 \subseteq \Theta_2$ with respect to a fixed likelihood $\lambda(X|\cdot)$, if there is a bijection $f : A_1 \to A_2$ such that, for all $S \subseteq A_1$, $\Pr_1[S|A_1] = \Pr_2[f(S)|A_2]$, and $\lambda(X|\theta) = \lambda(X|f(\theta))$, for all $\theta \in A_1$.

(81) *Fact:* For any pair of model framework parameter values ω_1, ω_2, $\Pr[X|\omega_1] = \frac{\Pr[A_1|\omega_1]}{\Pr[A_2|\omega_2]} \cdot \Pr[X|\omega_2]$.

(82) Returning to our particular case, in which the reduced model has $p^{(1)}_{0,1} = p^{(1)}_{1,1} = 0$: f can be the identity in this case, $A_2 = \Theta_1 = \Theta_2 = \{\langle 0,0\rangle, \langle 0,1\rangle, \langle 1,0\rangle, \langle 1,1\rangle\}$, and $A_1 = \{\langle 0,0\rangle, \langle 0,1\rangle\}$.

(83) *Right-hand ratio:* relative fit of the reduced model as compared to the full model (a function of the fits of the individual parameter vectors and their prior probabilities); can never favor the reduced model (must be at most one)

(84) *Left-hand ratio:* relative prior probability of the parameter vectors permitted under the reduced model, taken as a set. Since this is by definition one under the reduced model, this ratio can never favor the full model: it must be at least one

(85) This is a consequence of *monotonicity*: if $A \subseteq B$, then $\Pr[A] \leq \Pr[B]$, for any probability distribution; since $A_2 \subseteq \Theta_2$, $\Pr[A_2|\omega_2] \leq 1$

(86) More generally, for nested models ($A_1 = \Theta_1$, $A_2 \subseteq \Theta_2$):

(87) $$\frac{\Pr[\omega = 1|X]}{\Pr[\omega = 2|X]} = \frac{1}{\Pr[A_2|\omega = 2]} \cdot \frac{\int_{A_2} \lambda_{f(\theta)} p^{(2)}_{f(\theta)} \, \mathrm{d}f(\theta)}{\int_{A_2} \lambda_{f(\theta)} p^{(2)}_{f(\theta)} \, \mathrm{d}f(\theta) + \int_{\overline{A_2}} \lambda_t p^{(2)}_t \, \mathrm{d}t}$$

(88) In the most general case:

(89) $$\frac{\Pr[\omega = 1|X]}{\Pr[\omega = 2|X]} = \frac{\Pr[A_1|\omega = 1]}{\Pr[A_2|\omega = 2]} \cdot \frac{\int_{A_2} \lambda_{f(\theta)} p^{(2)}_{f(\theta)} \, \mathrm{d}f(\theta) + \frac{\Pr[A_2|\omega=2]}{\Pr[A_1|\omega=1]} \int_{\overline{A_1}} \lambda_s p^{(1)}_s \, \mathrm{d}s}{\int_{A_2} \lambda_{f(\theta)} p^{(2)}_{f(\theta)} \, \mathrm{d}\theta + \int_{\overline{A_2}} \lambda_t p^{(2)}_t \, \mathrm{d}t}$$

Appendix: How to construct other examples (e.g., OT)

(90) Not a lot of obvious places on OT grammars to hang the sorts of orders that give rise to BOR: like a P&P parameter space, a set of total constraint rankings is easily seen as "flat"

(91) *Possible soft spots:* stratified hierarchies (but generally not target states); phonotactic constraint induction, need to decide how fine-grained constraints should be; how much abstractness in the lexicon (but that would require an argument against lexicon optimization); constraint conjunction, assuming that this makes some induction required (not necessary in any case, but perhaps natural if self-conjunction is allowed, making the maximum constraint size unbounded)

Appendix: Size

(92) *Size:* For finite parameter spaces, we can also make essentially the same argument based on the size—if there are $N_1 < N_2$ parameters in the reduced versus full model, then the probability of each must be $\frac{1}{N_1} > \frac{1}{N_2}$—, but then we need to rely on a distributional assumption (uniform) rather than the more general MEP_2

(93) We could derive this from the following, which is also a weakening of (63).

(94) *The value of a sequence of rules is fully determined by the number of symbols in its minimal representation.*

(95) This should yield uniform distributions; we could then play off the fact that the set of grammars of size k is finite and smaller than the set of grammars of size $k+1$ (despite not being a subset).[2]

(96) However, we do not even need to assume (94). We can instead make the MEP assumption.

Appendix: Model evaluation versus comparison

(97) *Model comparison:* Infer a "model framework" hyperparameter explicitly

(98) *Model evaluation:* Do inference *as if* there were such a parameter: in Bayesian inference, can go straight for the parameter value by averaging over values of ω ("integrating out"), and the prior will be more concentrated on simple parameter values (compared to the "full" distribution, over all parameter values) just by virtue of incoporating a choice of models

[2] $\Pr[G|\omega_1] = K_1$ for all $G \subseteq \text{supp}\,[\cdot|\omega_1]$, $\Pr[G|\omega_2] = K_2$ for all $G \subseteq \text{supp}\,[\cdot|\omega_2]$, …, $\Pr[G|\omega_p] = K_p$ for all $G \subseteq \text{supp}\,[\cdot|\omega_p]$, …; the only way to get these conditional distributions to be constant is to make $\Pr[G|\omega_p] = \frac{1}{N_p}$. Suppose the support of the conditional distributions $\Pr[\cdot|\omega_p]$ is genuinely disjoint (that is, no grammar is considered to have both p and $q \neq p$ symbols in its minimal representation). Then $\Pr[G] = \Pr[G|\omega_p] \Pr[\omega_p]$ for the appropriate p.