ACL Lifetime Achievement Award

The Brain as a Statistical Inference Engine—and You Can Too*

Eugene Charniak†
Brown University

There are several possible templates for award talks.[1] The most common is an intellectual history—how I came to make all these wonderful discoveries. However, I am never completely happy with my work, as it seems a pale shadow of what I think it should have been. Thus I am picking a different model—things we all know but do not say out loud because we have no evidence to support them; and besides, making such bold claims sounds pretentious.

Thus I do not expect to say anything too novel here. I hope all my readers already know that the brain exploits statistics, and most of them suspect that we in statistical computational linguistics have something to say about how this works out in the case of language. My goal is therefore not to say anything you do not believe, but to cause you to believe it more passionately.

1. Evidence for Statistics

There is a growing body of evidence that our brains do their work via statistics. Here I present two studies: one classic, one new.

1.1 Lexical Acquisition in Infants

The classic work is that of Saffran and Newport (Saffran, Aslin, and Newport 1996) (S&N) on eight-month-olds' acquisition of lexical items. As is well known, speech is heard as a mostly unsegmented stream, thus raising the question of how children learn to segment it into words. What S&N show is that infants use statistical regularities in the input. When the stream is mid-word, there are fewer possible continuations than between words because of the uncertainty in the next word. More technically, the per-phoneme entropy is higher between words than within them. S&N show that eight-month-olds are capable of detecting such differences.

To do this S&N create an artificial "language" in which each "word" consists of three arbitrary phonemes, for example, *bidukapupadotigolabubidaku*.... So *biduka* and *pupado* are words, but the second two syllables of the first plus the first syllable of the second (*dukapu*) is not. All the words are played with no emphasis on any syllable and

* With apologies to Stephen Colbert.

† Department of Computer Science, Brown University, Box 1910, Providence, RI 02912.
E-mail: ec@cs.brown.edu.

1 I have written this paper in the first person to reflect its origins as the Lifetime Achievement Award talk at ACL-2011. It is not, however, based upon a transcript, since I can write better writing than I can speak. I also have included a few things I did not have time to say in the original version.

no difference in the spacing between syllables. (It is pretty boring, but the child is only subjected to two minutes of it.) After that the child is tested to see if he or she can distinguish real words from non-words.

To test, either a word or a non-word is played from one of two speakers. This is not done until the child is already looking at that speaker and the word is replayed until the child looks away. The children are expected to gaze longer at the speaker that is playing a novel (non-) word than for words that they have already heard.

Thus there are two testing conditions. In the first, the non-words are completely novel in the sense that the three-syllable combination did not occur in the two minutes of pretest training. On average, the children focus on the speaker 0.88 seconds longer for the novel words.

The second condition is more interesting. Here the non-words are made up of sound combinations that have in fact occurred on the tape, but relatively infrequently because they consist of pieces of two different words. Here the question is not a categorical one (Have I heard this combination or not?) but a statistical one: Is this a frequent combination or is it rare? Now the focus differential is 0.83 seconds. The conclusion is that children are indeed sensitive to the statistical differences.

1.2 What You See Where You Are Not Looking

Try the following test. Keep your gaze on the plus sign in the following example and try to identify the letters to its left and its right.

A + B R A C E

The "A" on the left is not too hard. The letters on the right are much harder, a phenomenon called "crowding." The work I am following here (Rosenholtz 2011) looks into this and related phenomena.

Obviously once we move our gaze around we have no problems with the letters. The center of the eye, the fovea, sends the brain a very detailed description. But elsewhere the brain gets many fewer bits of information. These bits have to "summarize" that piece of the image. The question that Rosenholtz (2011) looks at is what information these bits encode.

Suppose we want a 1,000-bit summary of the top left image in Figure 1. The other three images offer three possibilities. The top-right image simply down-samples the pixels. This is clearly not what the eye is doing. Going back to the crowding example, we cannot make out the letters on the right, but we can be pretty sure they are letters. Furthermore, we have little difficulty identifying the letter on the left, so it is not just down-sampled. The bottom-left image also down-samples, but on wavelets. It is little better, so we can ignore the fact that many of us don't know what wavelets are.

The bottom-right image is the most interesting. It assumes that the brain receives a basket of statistics found useful by the statistical vision community for summarizing images in general and textures in particular. The image shown here is a sample from the posterior of the statistics for the original (leftmost) image. Here it is reasonably clear that we are looking at letters, but we cannot be sure what they are, matching introspection in the crowding example.

If these are the statistics available to our brain, then looking foveally at such a reconstruction ought to be similar (equally difficult) to looking non-foveally at the original.

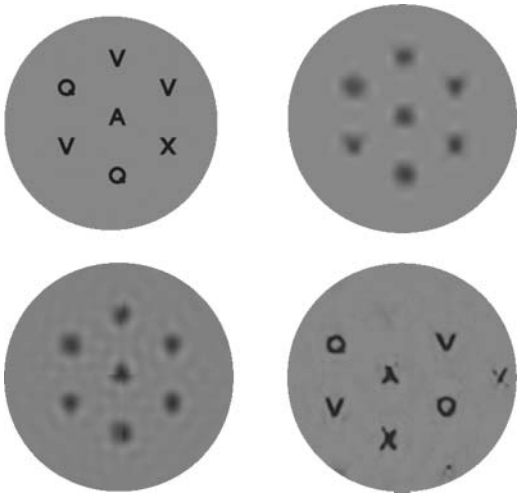

Figure 1
Simulations of the information available on non-foveated portions of an image.

The work in Rosenholtz (2011) shows that this seems to be true, again supporting the idea that brains manipulate statistical information.

2. Bayes' Law

Once one accepts that the brain is manipulating probabilities, it seems inevitable that the overriding equation governing this process is Bayes' Law:

$$\mathrm{P}(M \mid E) = \frac{\mathrm{P}(M)\mathrm{P}(E \mid M)}{\mathrm{P}(E)}$$

where M is the learned model of the world and E is the relevant evidence (our perceptions). We take Bayes' Law as our guide in the rest of this talk because there is so much to be learned from it. More specifically, let us use it in the following form:

$$\mathrm{P}(M \mid E) \propto \mathrm{P}(M)\mathrm{P}(E \mid M)$$

This is the case because $\mathrm{P}(E)$ acts as a linear scaling factor and thus can be ignored.

First, Bayes' Law divides responsibility between a prior over possible models ($\mathrm{P}(M)$) and a term that depends on the evidence, its posterior given the model. The war between rationalists and empiricists seems to have quieted of late, but it raged during much of my academic career, with the rationalists (typically followers of Chomsky) insisting that the brain is pre-wired for language (and presumably much of the rest) and the empiricists believing that our model of the world is determined primarily by the evidence of our senses. Bayes' Law calmly says that both play large roles, the exact balance to be determined in the future. The next section looks at the role of informative priors in current models of learning from sensory input.

Secondly, Bayes' Law says that evidence is incorporated via a generative model of the world ($\mathrm{P}(E \mid M)$). This is, of course, as opposed to a discriminative model ($\mathrm{P}(M \mid E)$). There is a lot to be said in favor of the latter. When both are possibilities, the general

wisdom has it that discriminative models work better on average. I certainly have that impression. But children have only the evidence of their senses to go by. Nobody reads them the Penn Treebank or any other training data early in their career. Thus generative models seem to be the only game in town.

Furthermore, this generative model has to be a very large one—one of the entire world insofar as the child knows it. For example, we can describe visual experiences in language, so it must include both vision and language, not to mention touch and smell. Thus it must be a *joint* model of all of these. Taking this to heart, I look in Section 4 at some work that uses joint modeling of multiple phenomena.

Lastly, Bayes' Law gives another clue, this time by what it omits—any notion of *how* any of this is done. This suggests, to me at least, that the inference mechanism itself is not learned, and if not, it must be innate. Or to put it another way, Darwinian selection has already done the learning for us. In the final section I will address what we can say about this mechanism.

One last point before moving on. What is the model that is to be learned? Bayes' Law does not, in fact, tell us. It only says that different models will be supported to differing degrees given the evidence available.

Various answers have been suggested. One group tells us that "true" Bayesians do not adopt any model, they integrate over them. That is, if you don't know the correct model, you should plan accordingly and not commit. This corresponds to the equation

$$\mathrm{P}(E) = \sum_{M} \mathrm{P}(M)\mathrm{P}(E \mid M)$$

Now integrating over all models makes a lot of sense, but to me it sounds a lot like telling people to change all their passwords every month—it is sound advice, but who will take the time?

One possibility is to adopt the most likely one, for example,

$$\arg\max_{M} \mathrm{P}(M)\mathrm{P}(E \mid M)$$

At first glance this would seem to be a no-brainer, but some worry that we could have a probability distribution something like that in Figure 2.

Here the most likely one, off on the right, has the very bad property that if we are only a little bit wrong we end up with a very bad model of the world, thus negatively impacting the probability that we will survive to have progeny. Better is to take the *average* model. This will put us somewhere in the middle of the large block on the left, relatively safe from catastrophic results from small errors.

My personal opinion is that this will turn out to be a non-problem. Given the right prior, the probability space over models will have one huge peak and not much else. Furthermore, as I discuss in the last section, our options on inference are going to

Figure 2
Projection on a line of models vs. probability for a bad case.

constrain us quite a bit, with an integration over comparatively few models coming out on top.

3. Informative Priors

Bayes' Law shows us how priors on possible world models should be combined with the evidence of our senses to guide our search for the correct one. In this section I give three examples of such priors, two in language, one in vision.

3.1 Priors in Word Segmentation

In section 1 we looked at the work in Saffran, Aslin, and Newport (1996) on infants' ability to divide a speech stream into individual "words." This problem has also been attacked using computational approaches, albeit with simplifying assumptions. We base our discussion on the work of Goldwater and Griffiths (2007), as it dramatically shows the need for *some* informative prior.

This, like most other computational linguistic work on word segmentation, considers an abstract model of the problem. A corpus of child-directed speech is translated into a simplified phoneme string by first transcribing the words, then for each word writing down its most common pronunciation. All spaces between sounds inside an utterance are removed, but the boundaries between utterances are kept. For example, *you want to see the book* would come out as *yuwanttusiD6bUk*. The output is to be the phoneme sequence with word boundaries restored. Although we are really dealing with phoneme sequences with spaces added, we will speak of dividing the speech stream into words.

A simple generative model for this task goes as follows:

```
For each utterance:
    Repeat until the "end of utterance" symbol is chosen
        Pick the next word according to P(w).
```

It is assumed here that P(w) is over possible vocabulary items plus a special "end-of-utterance" symbol.

If we have no prior on possible $P(w)$'s, then Bayes' Law reduces to finding the $M = P(w)$ that makes the data most likely, and this is easy to specify. It is simply a distribution with n "words," each one a single utterance. That is, it is a distribution that memorizes the training data. It is easy to see why this is so. If the model generalized at all, it would assign probability to a sequence that is not in the input, and thus the probability of observed training data must be less than that assigned by the memorization model.

In Goldwater and Griffiths (2007) the model was forced away from this solution by adopting a "sparse prior"—in this case a "Dirichlet" distribution. This says, in effect, prefer M's with as few different words as possible. Thus if the number of words is significantly smaller than the number of utterances, there is some pressure to prefer a word-based distribution over one for each utterance.

Unfortunately, the Dirichlet is a very weak prior, that is, the "pressure" is not very great. So in one solution to the problem the child's utterance comes out as youwant to see thebook. There is still a distinct tendency to merge words together.

Why is Dirichlet so weak? We want to evaluate the probability of our prior multiplied by a generative posterior for a particular model M. In our case M is just a

probability distribution over possible words. We assume for the sake of argument that we get this distribution from estimated integer counts over possible words. (In Section 5 we look at how Goldwater and Griffiths [2007] actually infers this distribution.) So our current guess is that we have seen the "word" *D6bUk* two times, and so forth. A simple maximum likelihood distribution would assign a probability to this word of $\frac{2}{L}$, where L is the total number of word tokens the model currently proposes in all of the utterances.

Of course, the maximum likelihood distribution is *called* the maximum likelihood distribution because for a given set of counts it assigns probabilities to make the probability of the data (the "likelihood") as high as possible. Therefore it will lead us to the memorization result.

The Dirichlet does something slightly different. Imagine that before we do the division $\frac{2}{L}$ we subtract $\frac{1}{2}$ from each nonzero word count. Then *D6bUk* has a probability $\frac{1.5}{L-0.5K}$, where K is the number of different word types (dictionary entries, as it were). This creates sparsity in the sense that entries that already have few counts tend to get still fewer because they are impacted more than words with high counts. So if we have two words, one with a count of 1 and one with a count of 1,000, and we have, say, a total of 10,000 word tokens and 1,000 word types, the first will go from a probability of 10^{-5} to approximately $0.53 \cdot 10^{-5}$ and the second will actually increase slightly (from 0.1 to 0.105). Thus words with small counts get crowded out, and the distribution becomes more sparse.

But the crowding is not all that strong, and in this model there are much stronger countervailing winds. The model assumes that words appear at random, and this is very far from the case. In the first utterance the phrase *the book* occurs much more often than the independent probabilities of *the* and *book* would suggest, so the model has an incentive to make them one word in order to capture a regularity it otherwise has no way to represent. Including bigram probabilities in the model would do the trick, and indeed, performance is greatly enhanced by this change. But this takes us beyond our current topic—priors.

3.2 Priors in Image Segmentation

We turn now to work by Sudderth and Jordan (2008) on segmenting an image into pieces that correspond to objects and background, as in Figure 3. To put the goal another way, each pixel in the image should be assigned to exactly one of a small number of groups, so that all the pixels in the same group exactly cover the places in the image corresponding to one type of object, be it people, trees, sky, and so on.

Any one pixel can, a priori, be in any of the different groups. But not all groupings are equally plausible in an everyday image. Consider Figure 4. Each row corresponds to random samples from a prior distribution over possible images, where each color corresponds to one pixel set. Although neither looks like scenes from the world around us, the upper row is more plausible than the lower one. The lower one corresponds to the so-called "Potts" prior: each pixel is a node in a Markov random field, and the Potts prior encourages neighboring points to be from the same region. Although these pixel arrays have such a property, there are all sorts of features that are not image-like. Single pixels are surrounded by other regions. If we divide the array into, say, tenths, each tenth probably has at least one pixel from all of the sets, and so forth.

The upper, more plausible "images" come from the prior illustrated in Figure 5. The process of randomly selecting from the prior works as follows: First generate a three-dimensional surface by picking a small number of points and placing a two-dimensional Gaussian over them. The surface point height corresponds to the sum of all the Gaussian

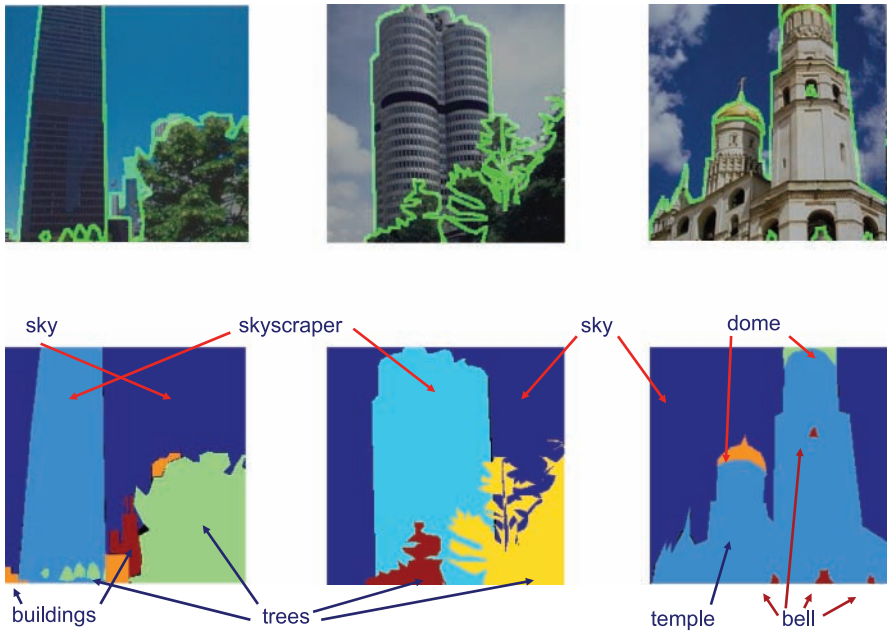

Figure 3
Several scenes segmented (by people) into their basic parts.

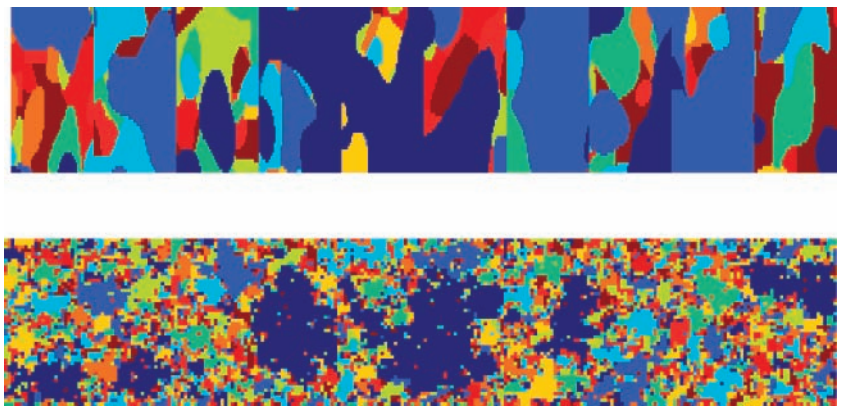

Figure 4
Samples from two image priors.

values at that point. The topmost image on the left is an example. Now locate a random plane parallel to the bottom of the image. All surface areas above the plane correspond to image region number 1. The dark blue area in the image on the right corresponds to a plane through the top left Gaussian with the yellow and red areas above the selected plane. We assign region 1 to be the areas most in the foreground.

We then repeat this process with new Gaussians and new planes (images 2 and 3 on the left) until we get the desired number of regions. The result is the random pixel array on the right.

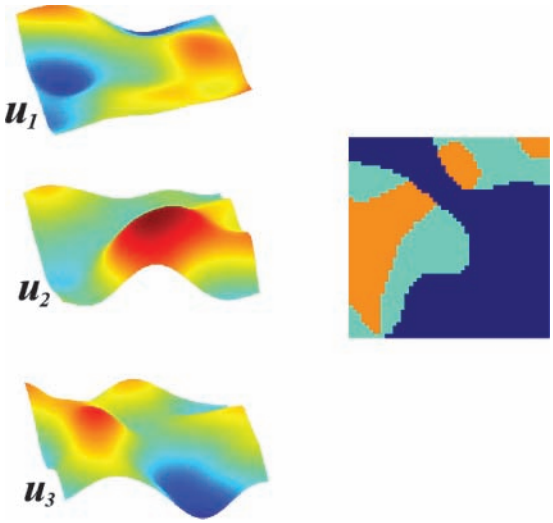

Figure 5
Image priors from sets of Gaussians.

Naturally, in image segmentation one does not generate random selections from the prior. But the depictions give the reader an idea of the biases placed on image reconstruction. A better way to think about the process would be to imagine that we try different reconstructions of the kinds of texture, lines, and so forth, that correspond to a single image region, and then present each of these region combinations to the prior for its "take" on things. It will strongly prefer blobby regions, with nearer ones tending to be convex. And, as we would hope, region segmentation works better with the more realistic prior.

3.3 Priors in Part-of-Speech Induction

Part-of-speech tagging is the process of labeling each word of a string with its correct **part of speech**, a category such as "plural noun" (NNS) or preposition (PP), that to some degree captures the syntactic properties of the word. The problem is not completely well defined insofar as there is no single "correct" set—part-of-speech sets for English range in size between 12 and $20 \cdot 12$. Nevertheless, words do fall into such categories and even as rough approximations they are useful as a first step in parsing. In unsupervised part-of-speech induction we want our program to infer a set of parts of speech. Because of their arbitrary properties, the size of this set is typically set in advance—for English typically 45, corresponding to the Penn Treebank set.

Since the early work of Merialdo (1994) we have known that this is a difficult problem, and in retrospect it is easy to see why. Most generative models work from the following generative story:

For each word w_i
  Generate t_i from t_{i-1} according to $P(t_i \mid t_{i-1})$
  Generate w_i from its tag according to $P(w_i \mid t_i)$.

where t_i is the ith (current) tag, t_{i-1} the previous, and w_i is the current word. With no prior over models, any such story is simply going to divide words into some number of classes to best raise the probability of the words. Merialdo uses **expectation maximization** (EM) to find his classes. Nothing in this set-up says anything about looking for syntactic properties, and it should therefore not be surprising that the classes found are as much semantic or phonological as syntactic. For example, a good way to raise probability is to split determiners into two categories, one including *a*, the second *an*. Similarly, for nouns create two classes, "starting with vowel" and "starting with consonant." This means that other less-used categories (foreign word or symbol) get omitted entirely. This is the sort of thing EM does. Words are assigned many tags (typically on the order of 10 to 20) and it evens out the tag assignments so most tags have the same number of words assigned to them.

There is, however, prior knowledge that is useful here. I am following here the work of Clark (2003). Clark first notes that, for the languages tried to date, most words have either just one tag or one tag that dominates all the others. As just remarked, this is a far cry from what EM does. So the first thing Clark (2003) does is to impose a one-tag-per-word-type constraint. This more closely models the true distribution than does the EM outcome. In the paper he notes that giving this up inevitably yields poorer results.

Secondly, he notes that, again contrary to EM's proclivities, tags have very skewed word counts. Some, typically content-word types like nouns, have very large numbers of word types, whereas grammatical types such as prepositions have very few. Putting this fact into the prior also helps significantly.

Note that using these facts (close to one tag per word and widely differing word-type counts per tag) in our prior amounts to positing linguistic universals. I am not an expert in world languages and thus anything I say about this is far from definitive, but as linguistic universals go, these look reasonable to me.

4. Joint Generative Modeling

Bayes' Law tells us that the proper way to connect the prior on a world model to our sensory input is via $P(M)P(E \mid M)$. We now consider the second term of this product, the generative model of the environment. Earlier we noted that this must be a joint model of vision and language because we can talk about what we see. Although work is beginning on such modeling, vision is (currently) a much harder problem, so joint work tends to help it and does little to help language researchers. However, there are benefits from combining different language problems into a single joint model, and in this section we consider two such cases.

4.1 Part-of-Speech Induction

In our discussion of priors for learning parts of speech we considered words to be atomic in the sense that we did not look at how they are built up from smaller parts (e.g., *playing* consists of *play* plus the progressive marker *ing*). This is the study of **morphology**.

English does not have a particularly rich morphology, especially compared to so-called free-word-order languages such as Russian, or Czech, but even in English the spellings of words can give big hints about their parts of speech. *Playing* in the Penn Treebank has the part of speech VBG, for progressive verb, and the *s* in *plays* indicates a present-tense third-person-singular verb (VBZ). Thus the tag induction program in

Clark (2003) actually computes the joint probability of the part of speech and the morphology. To a rough approximation, it uses the following generative model:

For each word w_i
  Generate t_i from t_{i-1} according to $P(t_i \mid t_{i-1})$
  Generate w_i from its tag according to $P(w_i \mid t_i)$.
    If we do not have a spelling for w_i, generate a sequence
    of characters $c_i, \ldots c_n$ from the single-letter model $P_{t_i}(c)$.

This substantially improves the model.

As noted in Clark (2003), however, this is not a true morphological model insofar as it does not model the *relations* among a word's morphological forms. For example, if we see "playing" and correctly identify it as a VBG, then the probability of "plays" as a VBZ should be higher than say "doors" (ignoring the probabilities of the other letters of the words). Morphology, however, is still difficult for our models as it requires learning multi-letter rule paradigms, and so far spelling by itself has worked quite well.

4.2 Joint Named-Entity Recognition and Co-Reference

Named-entity recognition assigns objects in a discourse (those that have names) to a small set of classes—prototypically, persons, places, and organizations. In another case of joint modeling, the Haghighi-Klein unsupervised co-reference model (Haghighi and Klein 2010) also learns named-entity recognition. (It can to some degree also learn entity types that are not realized by names, but we ignore that here.) Consider a simple co-reference example:

> Congressman Weiner said that his problems were all Facebook's fault. They should never have given him an account.

Suppose the program were able correctly to separate person names from those of organizations. This in turn would allow it to learn that people are referred to by *he*, *she*, and their variants, whereas organizations take *it* and *they*. That, plus the fact that entities with names are more likely to be antecedents of pronouns, would solve the example. (The program should be able to learn these facts about pronouns simply by looking at which pronouns are most likely to follow people/organizations. Naturally, there is no guarantee that the first pronoun following a person's name refers back to that individual, but the statistical signal is pretty strong.)

Conversely, suppose we have solved the co-reference problem. We could then pick up facts such as *Congressmen* is a good descriptor of one kind of entity (e.g., person). Or that people and organizations, but not places, can *say* things.

In practice the model distinguishes among entity types (e.g., person), entities (e.g., the particular congressman), and mentions (the word sequence *Congressman Weiner*). All entities have an entity type that provides a prior on the properties of the entity. So suppose we see a new entity that is described by the mention *Congressman Boehner*. The program will make a guess about entity type. If the program already "knows" that *congressman* is a legitimate descriptor for people but not other types, this will bias the program towards person.

Naturally, the program does not call these types "person," "place," and so on. Rather, the program is told in advance to expect some small number of entity types, say ten. If it works properly, we should see after the fact that the program has indeed

jointly learned not only co-reference, but the common types of individuals mentioned in the news articles.

Of course, the program cannot even begin to approach the accuracy of discriminative named-entity tagging. But from the perspective of this talk, that is besides the point. We are taking a child's view of the problem. Furthermore, eventually we want not just three or ten entity types, but a very detailed hierarchy (almost certainly an acyclic graph) of them. My guess is that such a graph can only be learned.

5. Inference

Bayes' Law says nothing about the mechanism used to infer a model of the world and from that I conclude that this mechanism is innate. Nevertheless, any model, no matter how abstracted from the particulars of neural implementation, must eventually confront the question of how these inferences are carried out.

For many of us, optimizing generative models has been virtually synonymous with expectation maximization. It is relatively efficient, and although it can be prone to local maxima, it often works well when started from uninformative initial states—the IBM Model 1 of machine-translation fame being an obvious case in point.

However, it has very serious drawbacks: It does not make room for a prior over models, it requires storing the training data for multiple iterations, and as its sine qua non, it requires the storage of all possible expectations on the data, which in many cases of interest is simply not feasible.

Consider again the problem of segmenting a sound stream into words. The string *yuwanttusiD6bUk* can be broken up into *yu want tusiD6b Uk*, *yuwant tu siD6bUk*, and so forth. To a first approximation the number grows exponentially in the length of the string and EM requires storing expectations counts for all of the possibilities.

This is why Goldwater and Griffiths (2007) use Gibbs sampling rather than EM. In Gibbs sampling we do not store the probabilities of all possibilities, but rather repeatedly go through data making somewhat random choices based upon the program's current "beliefs." For word segmentation this means we look at the point between two phonemes and ask, "Should we put a word boundary here?" We compute the probability for the two cases (two words here or just one) based upon the counts we have currently guessed. So if the "no-segment" case produces a word we have seen reasonably often, its probability would be higher than the product of the probabilities for the two-word case. Suppose it comes out 60–40. We then flip a 60–40 biased coin and put in a split if it is heads. Note that at any one time we need to store only counts for those words for which the program has at least one attestation, a number that grows only linearly in string length.

But Gibbs sampling, too, has serious drawbacks. Like EM, it is an iterative algorithm. As a psychological theory it would require the child to store all of her experiences and continually revise her beliefs about all of them. Another major problem with Gibbs sampling is its requirement that the probability model is "exchangeable." Each Gibbs iteration requires visiting all the decisions and (possibly) revising them. Being exchangeable means that if we had made the same decisions but in a different order (picking the sentences in reverse order, or first picking even-numbered places between phonemes, then odd), the probability after visiting every one exactly once would come out the same. Amazingly, the word-segmentation model is exchangeable.

But mathematical miracles do not grow on trees. Although the Haghighi-Klein reference model uses Gibbs sampling, their model is not in fact exchangeable, something

they had to work around. More generally, this requirement is ever less likely to be satisfied as our probability models become more complicated.

Or consider the options in inference mechanisms for statistical parsing. If we restrict ourselves to constituent parsing (as opposed to dependency parsing), the most used search algorithm is Cocke-Kasami-Younger (CKY). CKY takes a context-free grammar and uses the context-free property to reduce an exponential search problem to $O(n^3)$, where n is the length of the sentence. CKY is bottom–up, so for a sentence like *Dogs eat food* it would first find the noun phrases *Dogs* and *food*, then build up a verb phrase *eat food*, and only after completely processing the entire sentence would it link the subject *Dogs* to the verb phrase *eat food*.

But there are very strong reasons to believe that people do nothing like this. The strongest is the observation that various tests (as well as introspection) indicate that we are well aware that *Dog* is the subject of *eat* before the sentence is complete. (This is increasingly obvious as we make the sentence longer.) The second reason is more subtle but to me equally important, the context-free requirement. In a context-free grammar, saying that there is, say, a VP spanning positions 1 to 3 in our sentence is also saying that no knowledge from anywhere else in the sentence has any effect on what goes on inside this verb phrase, and vice versa. Putting it another way, every symbol in a CKY chart is a knowledge barrier between its outside and inside. We have managed to tame this somewhat by "state splitting," annotating the "VP" with other information, but to the degree that we do this we are exactly losing the performance that CKY should deliver.

We see a pattern here. We want large joint models, and we want to take advantage of them by using as much as possible of our experiential history (e.g., previous words of a sentence) to influence our expectations of what should come next and how our model of the world should adapt. But increasingly as we do this, various mathematical and algorithm shortcuts cease to work.

Actually, when you look at our requirements, there are not that many options. We want a mechanism that uses probabilities to guide the search for a model, it has to be able to incorporate prior knowledge, and it cannot depend on storing very much of our experiential history. Offhand this sounds like some sort of beam search with probability estimates as the search heuristic and this is, in fact, the solution that Brian Roark (Roark and Johnson 1999) and I (Charniak 2010) both adopted for parsing models that seem reasonable as models of human performance.

A standard example of such a search mechanism is **particle filtering**. It is somewhat like Gibbs sampling, except rather than repeatedly re-evaluating previous inputs it tries multiple evaluations immediately. In the limit, as the size of the beam goes to infinity, it is equivalent to Gibbs sampling but does not require exchangeability. But as John Maynard Keynes almost said, in the limit we are all dead. In practice we cannot afford that many particles (beam size). When discussing model selection, I commented that as far as I can see the only option is to "integrate" over a very small number of possible models. I said this because this is essentially what particle filtering is doing. Compared to EM, or Gibbs sampling, or CKY, this is quite inefficient. But it is not as though we have a lot of options here. It, or something quite like it, seems to be the best we can hope for.

6. Conclusion

I have argued that the brain is a statistical information processor and as such we should see it as operating according to Bayes' Law. On this basis we can see that learning depends both on a prior probability over models (our innate biases about how the world

works) and a joint generative model of the world. I also noted that Bayes' Law says nothing about inference mechanisms, and hence I assume that this mechanism is not learned—that it, too, is innate. I suggested that particle filtering, a beam search method based upon a probability evaluation function, seems a good bet.

Well, I said at the beginning that I would say little that most of my audience does not already believe.

But I also have had a subtext. I am not arguing that we as a field need to change our focus. Just the opposite. Our present focus has made statistical computational linguistics one of the most vibrant intellectual endeavors of our time. Rather, my argument is that we are already part-time cognitive scientists and together we are building a path to a new basis for the science of higher cognitive processes. I intend to follow that path. You can, too.

References

Charniak, Eugene. 2010. Top-down nearly-context-sensitive parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 674–683, Cambridge, MA.

Clark, Alexander. 2003. Combining distributional and morphological information for part of speech induction. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 59–66, Budapest.

Goldwater, Sharon and Tom Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751, Prague.

Haghighi, Aria and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393, Los Angeles, CA.

Merialdo, Bernard. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20:155–171.

Roark, Brian and Mark Johnson. 1999. Broad-coverage predictive parsing. Paper presented at the 12th Annual CUNY Conference on Human Sentence Processing, New York, NY.

Rosenholtz, Ruth. 2011. What your visual system sees where you are not looking. In *Proceedings of SPIE: Human Vision and Electronic Imaging*, pages 7865–7910, San Francisco, CA.

Saffran, J., R. Aslin, and E. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274:1926–1928.

Sudderth, Eric B. and Michael I. Jordan. 2008. Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Proceedings of Conference on Neural Information Processing Systems*, pages 1585–1592, Vancouver.