Class Notes

J.G. Makin

April 11, 2006

1 Caveat Emptor

If a lion could talk, we could not understand him.

Ludwig Wittgenstein, *Philosophical Investigations*

Medium rotation,
The shock of the new,
And a memo from Feldman saying,
'Everything is true.'

Dan Bejar, "The Sublimation Hour"

2 Introduction

These notes are intended as a supplement to the powerpoint slides, and not a replacement. There are many references to figures, none of which appear in this document, and most of which are extremely helpful if not absolutely necessary in understanding the material. So the reader is strongly encouraged to keep a copy of the slides at hand while reading through these notes; and to follow the external links, although they're usually of lesser importance. If there's a conflict between the content of these notes and the content of the slides, you should go with the latter—and you should also send an e-mail saying as much to makin+cs182@eecs.berkeley.edu. At the present moment, these notes are not complete.

3 Class # 3 (1/24/06)

3.1 Embryonic development

Nervous-system cells develop from the "ectoderm," one of the layers of the original embryo (the one from which skin cells also develop). There are two main types of these cells: neurons and glial cells.

Differentiation:

(a) *Where* the cell is born, with respect to chemical concentrations (e.g. SHH), determines what type they become: glial; motor and control; and interneurons, as the concentration decreases.

(b) *When* the cell is born determines where it ends up ("inside-out" arichitecture). Incidentally, radiation appears to (adversely) affect the migration of neural cells.

Axon guidance: Filopodia (little "feet" on the axon) follow the chemical gradients, bringing the axons to their proper targets. There are short-range as well as long range gradients.

Chemical gradients can be repulsive as well as attractive.

Connexions: Many axons connect to dendrites of "downstream" neurons. A winner-take-all mechanism prunes all but one axon. There can, however, be multiple growth cones from the single axon.

Activity dependent tuning: Chemical gradients lead neurons to their proper connecting sites, but there is "over-connection" (130%). These are then "pruned" in accord with neural activity.

4 Class #4 (1/26/06)

4.1 Neurobiological results

Sperry's experiment involved cutting the optic nerve in frogs and rotating the eye 180 degrees. The behavioral result: the frogs moved their heads *away* from food (flies). Conclusion: the optic nerve re-wired itself so that each area of the eye was connected to the same area of the optic tectum that it was prior to the surgery. Of course, now that the eye was rotated this was maladaptive. This re-growth indicated that the optic nerve wiring was genetically hard-wired.

On the other hand, Hubel and Wiesel's experiments demonstrate the importance of early post-natal development. The LGN in the thalamus is composed of layers, each of which receives input from a single eye. These layers are in turn connected to layer IV of the visual cortex. Now, cells that are connected to similar layers in the LGN tend to get "stacked together" in the visual cortex, forming columns of cells which are dominated by a single eye. However, if one eye is deprived of stimulus for a long period, esp. during weeks 1-9, then the ocular dominance columns never develop, no matter how long the eye is open later.

The fine-tuning need not be post-natal. In the womb, the infant moves its limbs, providing feedback (proprioceptive, etc.), and visual feedback is provided by systematic patterns of retinal stimulation. –Presumably this stimulation comes from the infant's own brain, so it is not a solution to the problem of limited genetic information: in some sense, this information had to be encoded, presumably in the genetic code. Rather, it is a solution to the problem of an infant's needing to see very soon after its birth. Similarly, colts are born able to walk.

Another example of pre-natal tuning is the auditory stimulation provided by the mother's voice, which includes language information: newborns have been shown to be respond preferentially to the sounds of their native language (this presumably is true for white people born in China as well as etc....)

There is of course an information-theoretic issue (cf. the hw ass't): our genes can only encode so much information; the rest of the constraints must be supplied by environment. That's one reason why the stunting of the formation of ocular dominance columns should not be surprising.

Having stressed genetic pre-wiring and pre-natal/post-natal fine tuning, we should mention that the brain is nevertheless plastic in many ways later in life. Recovery from brain injury (esp. stroke) and even the elimination of phantom limb are examples of this phenomenon. No one knows exactly how this happens, but it may be that redundant connections which had previously been subjected to inhibition are now "released" to carry information in ways they didn't before.

4.2 Introduction to McCulloch-Pitts Neurons

We now move on to some computational abstractions of the neuron, and the computational models that are based on it. What features do we want to capture?

- conductivity delays??
- discrete firing (on or off)? or maybe real numbers for firing rates?
- weighted inputs
- transformation of inputs to outputs ("activation-output relation")

The McCulloch-Pitts neuron is one such abstract computational model. Outputs are often discrete, delays are neglected, input is the inner product of a set of discrete inputs with a weight vector, and the activation-output relation may be chosen from any of a number of function types, typically: linear, threshold, sigmoid, or radial-basis. Note that the output is real-valued for all of these except the threshold function, though input to the whole network is still generally discrete. This means that the network maps from binary-vector space to the interval [0,1]. MP neurons also often have a *bias input*, which is essentially just a scalar offset of the inner product (so the function which computes the weighted input—which is not, it should be stressed, the activation-output relation—is affine rather than linear in the input space).

Analogies between the various parts of the neuron and the computational abstraction were made, and then the MP neuron was explored in some detail. Additional modifications were also mentioned: the input could be sigma-pi or cubic rather than linear, output could be made stochastic, etc.

Using only threshold activation functions, a neural network of MP neurons can separate input space into regions only with *hyperplanes*, which is a line for a two-input system and a plane for a three input system. Since a line cannot segregate the "on" pattern of (e.g.) an XOR from its "off" pattern, the XOR can't be computed w/this type of network.

5 Class #5 (1/31/06)

5.1 The McCulloch-Pitts Neuron and "triangle nodes"

McCulloch and Pitts invented an abstract model of neurons which could be used to compute logical functions. This model, the McCulloch-Pitts neuron, became the basis for artificial neural networks (ANNs). The nomenclature is, frankly, not terribly important, but ANNs with *threshold* activations are called MP neurons, and networks which consist of just one of these neurons are called *perceptrons*. Other activation functions are often used as well: commonly the radial-basis function and the sigmoid function.

A single-layer network with only threshold activation functions can compute all and only linearly-seperable functions. A two-layer network can compute any convex function, and with three layers (i.e. two hidden layers) we can compute whatever we want. Support vector machines and other machine-learning techniques are in some sense generalizations of these networks.

Triangle nodes will be an important element of our computational models, so it is worth showing that they can be implemented with our neuron models (see slides). Triangle nodes are our explanation for the binding of features of objects to values. (Artificial) neural networks are also the basis for *spreading activation*, which is an important part of our story on the neural/computational basis of thought (more on this later).

5.2 Psycholinguistic evidence for a global constraint theory

One hypothesis about our information processing is that it's done serially, with bits of information processed one at a time (the "modular view"). An alternative (which we lean towards) is that all information immediately imposes constraints on the processing. One experiment which supports the latter hypothesis is the "They all rose" priming experiment. The task: identify if the string x (e.g. *stood* or *flower*, or *gavigai*), which is presented on a monitor, is a real word. Simultaneous with the appearance of the word on the monitor, the subject hears a sentence containing a different, though related, word over headphones. This second word is either directly related to the word on the screen (as, e.g., the word *rose* in the sentence "They all rose" is directly related to the visual stimulus word *stood*); or it is used homonymously to a related word (as is, e.g., the word *rose* in "They all rose," where the visual stimulus is the word *flower*; the meaning of *rose* in this sentence is unrelated to flowers, but its homonym *rose* is). The auditory stimulus is either offset by 200 ms or not at all. The result:

- No offset: Priming effect for both *stood* and *flower*, i.e. the subject was quicker at recognizing them as real words
- 200 ms offset: Priming effects only for the appropriate sense ("stood")

Our conclusion is that there are connections in the brain between the neural representations of the (sound of the) word *rose* and the representation of the concepts of *stood* and *flower*, as you might expect. So *ceteris paribus*, activation of the former spreads to activate the latter two, and this is exactly what we see when there is no offset between presentation of the visual and auditory stimuli. However, we also hypothesize that the representations of *stood* and *flower* compete for activation, meaning that they mutually inhibit one another. If one of these two "nodes" (you can think of these representors as somewhere on the order of ten neurons) is elsewise activated, it will succeed in inhibiting its competitor. This is what we think happens in the case with the 200 ms offset: the context of the sentence provides extra activation to the representation of the appropriate sense of the word, which allows it to inhibit the inappropriate sense(s). Thus in the example given, comprehension of the sentence inhibited the node representing *flower* while activating the node representing *stood*, so subjects were quicker only at recognizing the latter as a real word.

Here's another experiment: Allopenna, Magnuson, & Tanenhaus used an eye-tracking device to determine if rhymes compete for activation. The idea is that the pictures which correspond to words which rhyme with the actual target word will receive greater eye attention than those that do not.

The cohort theory (Marlsen-Wilson) claims that onset similarity of words will provide the primary basis for eye competition. Neighborood Activation Model (Luce): Global similarity is primary, so rhymes provide the strongest such basis. TRACE (McClelland & Elman): Global similarity is important, but the incremental (serial) nature of speech constrains the result; so both cohorts and rhymes compete, but with different time courses. This last was vindicated by the experiment. How? Well a connetionist model based on this idea predicted time courses which matched the eye-tracking results very well. This again suggests that all constraints are considered, and that they begin constraining as soon as they are available.

6 Class #6 (2/2/06)

6.1 Techniques for electrophysiological investigations

Functional Magnetic Resonance Imaging (fMRI) measures blood flow (more precisely, the "hæmodynamic response"). It does so by measuring the magnetic response of the blood, which varies according to oxygenation (this variance is manifested in the blood oxygenation-level dependent [BOLD] contrast.) Since using part of your brain draws blood to it, just like moving a muscle does, the BOLD signal is a good indicator of what parts of your brain are involved in certain tasks.

It is not, however, conclusive, since brain areas showing increased blood flow are only a superset of the brain areas being used. (That is, we are fairly safe in assuming that any brain region which is active in processing a task will show increased blood flow, but other areas may show increased blood flow as well; so the increased BOLD signal is a necessary but not sufficient condition for relevance to the current task.)

Another shortcoming of fMRI techniques is the four to give second delay between the increase in neural activation and the changing BOLD signal. It's not that the signal appears all the same, just delayed by five seconds; rather, it is in effect integrated over that time scale, so precise temporal information is lost.

On the plus side, though, fMRI has (relatively) good spatial (as opposed to temporal) resolution, on the order of millimeters. That is, the fMRI images the brain into 2-4 mm^3 volume units, called *voxels*, on analogy with "pixels."

We also discussed an technique for determining brain function called transcranial magnetic stimulation, or TMS, in some detail. Here's how it works: a current is passed through a large coil of wire, which generates a magnetic field (recall Maxwell's equations from your physics or EE classes). The shape of the coil determines the shape of the magnetic field. The coil is placed next to the skull, and the current switched on and off in a precise way to control the magnetic field. The magnetic field in turn induces current flow in the brain (again recall how this works from Maxwell's equations).

From the investigator's perspective, the point is to have the subject try to perform certain tasks, and then to zap parts of the brain that are suspected to be involved in the performance of that task. If performance degrades or fails, the portion of the brain which has been stimulated is assumed to play a role in performance of that task.

TMS has also been used to investigate the motor cortex. Have the subject watch something that we suspect will activate the motor cortex. Generally speaking, this stimulation will not cause the subject to actually move his muscles, but it'll be close; his muscles will be just below the threshold for movement. Now, if we zap his motor areas with TMS while he's watching the stimuli, we can push him over the threshold and induce muscle twitches in the areas that were stimulated. So it's a way of figuring out exactly which motor areas were activated by the stimulus.

TMS can also be used in conjunction with fMRI. Since, as we said before, increased BOLD signals are necessary but not sufficient for the relevance of that brain area to the task at hand, we use TMS to disable this or that region seen to be activated in the fMRI, and see if the task can still be performed. This allows us to get that slippery sufficient condition. (That is: an increased BOLD signal in conjunction with an effective TMS inhibitory legioning in the same area is both necessary and sufficient for that area being involved in a particular task).

TMS is sometimes used purely as a stimulation technique, as well, as a form of treatment for epilepsy and for depression. If this sounds bizarre to you, gentle reader, you are not alone.

6.2 Neurological experiments

In 1996, Gallese and his colleages discoved neurons in the premotor area F5 which fired when the monkey either performed an action or watched someone else perform it—and not to other, similar hand actions. So, e.g., if there was no object present but the investigator performed the same (e.g.) grasping motion, the neuron didn't fire. Moreover, some of these neurons fired irrespective of the effector used (left hand, right hand, mouth). They were dubbed, for obvious reasons, "mirror neurons."

The data suggest that these macaques have "abstract" representations of goal-oriented actions, i.e. abstracted from who in fact performs it. This will play a major role in our theory about the neural basis of language. Do humans have a similar system? Well we can't do thorough single-cell recording in humans, but fMRI data from Buccino et al do suggest the following:

- Watching the performance of an action with or without an object produces activation in premotor areas. The activation sometimes appears weaker when the object *is* present.
- In the "with-object" cases, parietal areas are also activated.
- Both the premotor and parietal areas which are activated by observing actions are somatotopically organized; that is, which region is activated depends on which effector—hand, leg, or mouth—was used in the observed action.

We conclude that the frontal areas (i.e. in the premotor cortex) encode the goal of the action, which may be deduced regardless of whether the object is present and simply if the action is voluntary. (Buccino claims that the action must be voluntary to elicit activation in the observer's premotor cortex, though that's not clear from his study. I take it this has been established eslewhere.) The parietal areas, on the other hand, encode something more specific, something like the effect of the action (and hence only fire if the object is present).

Mirror neurons have been discovered more recently which are invariant to modality—or anyway certain modalities. In particular, the sounds or sights of certain events (ripping paper, breaking a peanut) activated these neurons. Moreover, if the monkey executed the task, say, of breaking a peanut, the same neuron fired. On the other hand, non-action related sounds (e.g. white noise) actually decreased firing rate. Furthermore, the neurons were task-discriminative: some fired for breaking peanuts, others for paper ripping, but not for both.

This suggests another form of abstraction, not just over who performs the actions, but over how the action is perceived (sight or sound). We'd like to conclude (and for you, gentle reader, to conclude) that these neurons are the representation of specific, action-oriented tasks. Most curiously of all, perhaps, the neurons were found in the monkey homologue of Broca's area—a human language area. (Here's the original Science article: http://www.sciencemag.org/cgi/content/full/297/5582/846)

Again we ask if there are analogues in humans. Similar to the TMS motor experiment described earlier, an acoustic stimulation test was performed on humans. While listening to auditory stimuli, the related motor area was zapped using TMS. Were the MEPs (motor evoked potentials) in the relevant effector (hand vs. foot) greater when the related auditory stimulus (typing or tearing paper vs. walking) was presented? Yes, which suggests that those motor areas are activated merely by listening to the sound of an action performed with the part of the body that that area controls!

This is a lot of data to digest, but here's the big idea. Language use has its basis in the ability to represent things abstractly, and we have found the neural basis for abstract action representations in monkeys. We'll suggest later that *all* of our linguistic abilities are rooted in our sensorimotor and affective apparatuses; that is, all of our representational (and inferential) capacities piggyback on our sensorimotor and affective (emotional) representational (and inferential) capacities. Evolutionarily speaking, it's not to hard to imagine goal oriented actions leading to pantomiming without the goal present, leading to an abstraction of the pantomime into the spoken word.

This is consonant with another notion we want to convince you of, namely that our semantics, i.e the meaning of words, is *embodied*, or in other words that they are organically connected to our getting around in the world, and therefore importantly to our goal-oriented actions.

Here's a way of putting the claim as a testable hypothesis: The same brain areas that process sensorimotor experiences also process the semantics related to that experience. An experiment was devised to test this hypothesis. Very simply, subjects were asked both to observe hand/mouth/foot actions and to read sentences that described these actions, and fMRI was used to see if the same regions were activated in both cases. The results are not conclusive, but this much can be said: Premotor areas that were activated most in the observation of hand or mouth (though not foot?) actions were also activated the most in reading phrases about (literal) hand or mouth movement—though these correlations only appeared in the left hemisphere (not surprisingly, since it is the language-processing hemisphere, broadly speaking). Why not the foot? one wonders. And will this work for metaphorical actions sentences? The theory we're pitching in this class (henceforth, the "party line") is that the same results will indeed be found for metaphorical sentences, since the party line is that *all* language processing uses our sensorimotor and affective circuits. (We may soften this strong claim later on, but for now take it as a good summary of the party line.)

7 Class #7 (2/7/06)

Memory can be divided into declarative (either episodic or semantic) and non-declarative (i.e. "procedural"). We care about these distinctions because we'd like to know how the brain learns each of these kinds of thing (episodes, facts, and skills, respectively), and there's reason to believe that different mechanisms are at work. In particular, we note that: declarative learning can be "one-shot," whereas non-declarative learning usually takes *training* (which from our computational perspective suggests ANNs).

Here are some different types of learning from a computational perspective:

1. Hebbian learning
    - coincidence learning
    - "neurons that fire together wire together"
    - Long-term potentiation (LTP) is the biological mechanism
    - LTP takes place at two different time scales:
        - times scale: hours-days
        - "Broadly, [early LTP] produces a potentiation of a few hours duration. It does so by making the postsynaptic side of the synapse more sensitive to

glutamate by adding additional AMPA receptors into the postsynaptic membrane." (from the wikipedia)

    - More detailed: Repeated application of presynaptic spiking (i.e. high frequency spiking) causes NMDA receptors in the post-synaptic cell to release the magnesium ions which blocked them. AMPA receptors then get activated by the newly-formed calcium-based substances in the post synaptic cell, making it more sensitive to input from the pre-synaptic cell. The excitation necessary to bring about the potentiation is a 100 Hz signal applied over the course of a second.
    - "Conversely, L-LTP [late LTP] results in a pronounced strengthening of the postsynaptic response largely through the *synthesis of new proteins*." (from the Wikipedia, emphasis mine)
    - Three spike trains, spaced apart by ten minutes, can set off late LTP.

- Winner-take-all computational mechanisms use Hebbian weight modification.
- Episodic learning is thought to be based on LTP.
- We also looked at a way that a Hebbian-learning network could acquire representations of letter (or phoneme) pairs (see slides).

2. recruitment learning
    - one trial
    - based on LTP, perhaps??
3. supervised correction
    - the mechanism is presented with the correct output
    - e.g. backprop
4. reinforcement learning
    - delayed reward
5. unsupervised
    - abstraction of similar features

We'll revisit these other (2-4) kinds of learning in greater detail in later lectures.

8 Class #8 (2/9/06)

Hebb's rule is powerful, and provides for lots of association-type learning. However, it's not really flexible enough to perform certain kinds of learning tasks. Let's consider a supervised-learning technique [that's (3) from class #7] called

8.1 Backpropagation

Backprop requires a continuous, differentiable activation function, which rules out the threshold function. So we use the sigmoid. A full derivation of the backprop algorithm appears in the handout, *which you should consult at this point!*

9 Class #9 (2/14/06)

9.1 Artificial neural networks: wrap-up

We once again talked about backprop, this time also giving some justification for the momentum term. Here's one way of thinking about it: A "true" gradient descent algorithm would take infinitesimal steps, so in fact the learning rate is like a short-cut. It lets us skip down the hill faster. But this will be a problem on "jagged" error surfaces, where we'll keep bouncing back and forth over narrow valleys, changing direction (sign) of weight change every time we hit the opposite wall. The solution is to add a term proportional to the *previous* weight change, so that we'll have an increased tendency to continue in the direction we just moved in.

What's the "representational power" of ANNs using the sigmoid activation function? Well a single-layer network (no hidden layer) can compute any linearly-separable function; two-layer networks (one hidden layer) can compute any convex functions; and three-layer networks can compute (basically) any function.

Another issue with ANNs is the balance between overfitting and ability to learn anything at all. In designing an ANN, we have to steer between these two, generally trading off one for the other. This decision manifests itself in (a) number of hidden nodes (more means more likely to overfit, less means less likely to learn) and (b) number of training steps. For the latter, we can use this trick: hold out some of the training data as "validation data"; a good place to stop training is when the network starts fitting the validation data worse (even though it continues to fit the training data better).

Finally, we talked about extensions to feedforward nets: Jordan nets, Elman nets, and recurrent nets. We didn't go into great detail on them but if you're curious and want to know more, they're all over the Internets, so look them up.

9.2 Recruitment learning

We now change gears and talk about (2) recruitment learning, one of the five types mooted in Class #7. Suppose we want to link two distant neurons or groups of neurons which each represent some concept. I.e., we want to link these concepts. What are the chances that these nodes are connected via some neurons? Answer:

$$P(p=0) = (1-F)^{B^k} \tag{1}$$

where $P(p=0)$ is the probability of that there are zero paths between two nodes that are k layers apart, given that each node branches out to B nodes in the next layer and that there are N nodes per layer. The branching factor, F, is just B/N. (Note that this is formula provides only a lower bound, since it assumes that no two nodes in the one layer output to the same node in the next, i.e. that the mapping is "injective." This is not necessarily true, but the true answer is close enough.)

So, e.g.: say we want to remember a fact like "Srini got his master's at Stanford." Now we may already have representations for Srini (I certainly do), Stanford, and of the relationship *received his master's degrees at* (R, for short); what we need is a way to link these together (note that we're obviously simplifying a bit here). That's where recruitment learning comes in: If the neural representations of *Srini*, *Stanford*, and R are sufficiently spaced (and the chances are good that they are), then there is almost certainly a path connecting them. We can now "recruit" one of these neurons (or a few of them) to represent our fact. This might happen along the lines of Hebbian learning: the activation of the orginal three representations activates adjacent neurons; the ones with the most activation are (we suppose) the ones that are connected to all three and these are the ones that are recruited to represent the relationship. We call nodes that serve this purpose "triangle nodes," for obvious reasons. Note that after learning, the nodes can be used to recall facts; so if, e.g., we wonder where Srini got his master's degree, we activate two of the three units connected by the triangle node, and via the latter, activation spreads to the third unit, that representing "Stanford." That's a rough sketch of our mechanism for learning and recalling such facts.

10 Class #10 (2/16/06)

This lecture was on color vision. The reason why we care about color is that it illustrates nicely some ways in which language use and language processing are tied to biological facts about our eyes and our brains.

10.1 The physiology of vision

The visible spectrum lies between about 400 and 700 nm.

- Hue = wavelength (or frequency; take your pick)
- Saturation = variance of wavelength distribution
- Brightness = area under the distribution

Rods are highly sensitive but just to light and darkness, not to color; cones require more light, are less sensitive, detect color. (This is why the world looks colorless at night.)

There are in fact three different types of cones, which are responsive over three different ranges of wavelengths. The ranges have peaks at:

- 437 nm (short/blue cones)
- 498 nm (rod)

- 533 nm (medium/green cones?)
- 564 nm (green/red cones)

“Computations” are done by the various layers of the retina. For instance, center/surround cells perform edge detection via lateral inhibition. Colors beyond those picked out by the three types of cones are computed by taking the difference of ”opponent processes”; e.g.,

- R+/G- = L - M
- G+/R- = M - L
- B+/Y- = S - (M+L)
- Y+/B- = (M+L) - S
- W+/Bk- = S+M+L
- Bk+/W- = -(S+M+L)

This is also done via lateral inhibition.

We also talked briefly about color constancy. We know that:
(Illumination spectrum) x (reflectance spectrum) = (luminance spectrum), i.e.
(# photons emitted) x (% photons absorbed) = (# photons perceived)

But how do we decide that the color is “the same” when the luminance spectrum is different? I.e., we can perceive constancy in reflectance despite differences in illumination. This is actually but one of a number of vision problems where input stimuli underdetermine the result. (For putative solutions to these problems, you can check out Bruno Olshausen’s work, `http://redwood.berkeley.edu/bruno/research/`.)

10.2 Color Naming

1. Fact: there are at most 6 “primary” color terms, where primary means

   - non-compound words;
   - colors that can apply to anything (so not, e.g., “blonde” or “roan”);
   - frequently used (not “mauve”); and
   - words which refer primarily to colors (not “lime”).

   This is true cross-linguistically—though there are some “missing” colors in some languages. That is, some languages group two or more of the aforementioned colors together into larger sets, and we consider these to be primary as well.

2. These color terms cut up “color space” the “same way,” in the sense that cross-linguistic overlap occurred only as subsets. I.e., boundaries didn’t cross each other.

3. The centers of these groups corresponded to the cones' peaks, i.e. at "focal colors."

4. There is a typical development sequence for differentiation of colors as a culture "evolves" (see slides)

5. Focal colors are more easily remembered; new color categories centered on the focal colors are more easily learned; and focal colors are categorized more quickly than peripheral ones.

6. In addition to the six primary colors (red, green, blue, yellow, black, and white), there are six "derived" colors (pink, gray, orange, purple, brown, and light-blue [in Russian]), which can be thought of us "fuzzy-ands" of various of the primary colors. We can think of the composite colors of more "primitive" cultures as "fuzzy-ors" of the primary colors. (We'll talk more about fuzzy logic later in the semester. In the meantime, some info can be found here: `http://www.seattlerobotics.org/encoder/mar98/fuz/flindex.html`.)

11 Class #11 (2/21/06)

11.1 Categories & Concepts

Concepts are (usually?) universal, stable, internally structured, compositional, inferential, relational, meaningful, and not tied to specific word forms. So we claim. [In fact, one might suspect that"Concepts are universal" is either trivially true or wrong....]

Here's an (invidious) contrast we'd like to strike up about for definitions of concepts:

(a) "Traditional" theory: concepts are abstract, disembodied, separate from perception and action, and human-specific.

(b) The NTL party line: human concepts are embodied in our sensory-motor, emotional, and social cognition capacities, and to the extent that our near relatives (e.g. primates) have these latter capacities, they have concepts.

And there's a parallel distinction about categories:

(a) Classical: Categories have clear boundaries, essential properties (necessary & sufficient conditions).

(b) Prototype theory: categories form around exemplars from experience. The co-occurrence of properties establishes category boundaries, but they aren't hard and fast. Members of a category need not all share the same properties.

We are pushing (b) in this class.

11.2 Basic-level categories

We can distinguish between basic, superordinate, and subordinate categories, where the latter two are defined in relation to the basic ones in the obvious way. What then is the distinguishing mark of the basic-level category? Answer:

- perceptual: overall shape, single mental image, fast identification;
- functional: we have "motor programs" for interacting with members of these BLC;
- linguistic: BLC members have the shortest words, are learned earlier, entered the lexicon earliest; and
- knowledge organization: most attributes are stored at this level.

Of course, to be consistent we have to claim that these criteria are not necessary or sufficient (i.e. not, in a word, "criterial," in the philosophical sense.)

11.3 Internal Structure of categories

We can think about categories in terms of their "internal structure" as well as their relations to other categories.

- Radial categories: there is a central case, and peripheral cases share features in common w/the central case but not necessarily the w/each other (e.g., our concept of "mother." (So we visualize the peripheral cases being connected by radii to the central category, hence the name.)
- Family resemblance: Some members of the category may resemble others in a certain way, and those latter resemble still other members in other ways; in the same way that I may resemble my brother in some features and he my sister in other ways, but I may not resemble my sister in those ways (or even any at all). These category relations are those known as "family resemblances" (coined by the phisopher Wittgenstein).)
- Prototype-based (e.g. sparrow for bird)
- Essentially-contested (e.g. democracy)
- Ad hoc (e.g. things to take on a picnic)

Concepts can become the prototypes for their categories by

- being a central subcategory, so that other relate to it (e.g. walk; amble, swagger, saunter);
- having "essential" characteristics: certain qualities almost always show up in folk definitions (e.g. birds have feathers & beaks, lay eggs);

- being a typical case: we are most familiar with one example (sparrow for birds);

- being the ideal/anti-ideal (notice that these might in fact not be very common, unlike stereotype or salient examplars), e.g. ideal vacation;

- stereotype, e.g. we may have a stereotypical notion of what a frat boy is like;

- salient exemplar, e.g. 9/11 for act of terrorism; and

- generators, e.g. the notion of finite-dimensional vector spaces was generated by abstracting properties of $\mathbb{R}^n$, so the latter, by generating the former, has become a prototypical finite-dimensional vector space.)

11.4 Brain Studies

Here's a PET study that provided evidence for neural correlates of category structures. Subjects looked at (1) pictures of animals, (2) pictures of tools, (3) noise patterns, and (4) novel nonsense objects. They were asked to silently name the former two to themselves; the latter two were used as baselines (i.e. the responses produced by these were subtracted from the responses for the stimuli of interest, in order to determine the differential activation of tools vs. animals). The stimuli were fixed for 180 ms apiece, followed by a fixation cross of 1820 ms. And indeed, although animals and tools both activated the ventral temporal lobe region, (naming) tool pictures differentially activated the anterior cingulate cortex ACC, pre-motor and left middle temporal region—which is known to be related to the processing of action words! Naming animal pictures differentially activated the left medial occipital lobe, which is involved in early visual processing.

Another brain study, this one using fMRI, set out to show that reading action words differentially activates the pre-motor area, and that the responses are *somatopic*, that is that which portion of the pre-motor area is activated depends on what body part is involved in the given action. Why suppose this? Well we already know from action studies that face, arm, and leg representations are somatotopically oganized in the motor and premotor cortex. If words as well as actions elicit (the same) somatotopic activations, then we have reason to believe that these motor areas play a role in language processing. This is one of the central claims in this class. And in fact, the study showed just that.

What conclusions can we draw from these studies about concepts and category structure? Both studies provide evidence against the traditional notion of concepts as being disembodied; on the contrary, certain concepts at least have a neural basis in our motor system. Second, it suggests that categories (e.g. animals and tools) are represented by distrbuted circuits in the brain, circuits which take part in sensory and motor tasks. Put more concretely, our categories for animals are represented at least in part by neurons in visual areas; and our categories for tools are represented at least in part by motor circuits, presumably for interacting with the relevant tool. Our categories for actions are organized according to, *inter alia*, the effector used in each; and represented by, *inter alia*, distributed motor circuits.

12 Class #12 (2/23/06)

12.1 Image Schemata

In this class, we talked about the way certain concepts are embodied in our image schemata (though, frankly, our schemata are multi-modal). In English, prepositions encode a lot of our spatial relations, and hence our understanding of them is image-schematic.

12.2 Cross-linguistic variation

There is some limit on the extent of schematic spatial distinctions. Most (all?) cultures distinguish between:

1. figure & ground,
2. relative orienations (and other geometric facts),
3. contact and its absence, and
4. force dynamics (independent of spatial distinctions).

However, we (students of psychology) must distinguish between:

(a) the image schemata associated with a particular word (e.g. “on”), which groupings *vary* across languages; and

(b) the “primitive” image schemata, which appear in various languages (though in various groupings).

In class we saw examples of the way prepositions in different languages cut up semantic space differently, while at the same time (apparently) making use of the same underlying image schemata—making use of them, i.e., in different ways.

Here’s a pithy way of summing up a weak form of the embodiment-of-language-and-thought claim: “Regularities in our perceptual, motor, and cognitive systems structure our experiences and interactions with the world.” There are also regularities in our environment, and in the way all people interact with it. Image schemata are an example of those regularities in our perceptual and cognitive capacities.

12.3 Spatial Image Schemata

1. *Trajector/landmark*: trajectors are located w/respect to landmarks; the latter are often fixed whereas the former are not. This schema manifests itself in linguistic asymmetries, e.g.
   - ?The post is next to the skateboard.
   - ?The table is under the cup.

2. *Boundaries & bounded regions*

3. *topological relations*: Separation, contact, inclusion & overlap, encirclement)

4. *Orientation*: Inherent (left/right/front/back), viewer perspective, external (bayside/hillside), absolute (N/E/S/W)

.......[Had to leave early]

13 Class #13 (2/28/06)

Terry Regier set out to build a computer program which could acquire a portion of the language; in particular, to learn spatial relations by pairing images which represented these relations to the words that encode them. Here are some issues that Regier had to deal with:

1. Recall from last lecture that spatial-relation words from various languages group image schemata together in inconsistent ways. (Contrast this with Kay's discoveries about colors. Languages with different color words still cut up color space either (a) the same way, or (b) in supersets or subsets of (say) English. Boundaries for colors in one language never cross boundaries of colors in other languages. This is *not* true for spatial relations.)

2. Context of objects which are neither trajector nor landmark also determine word usage. Recall from lecture the two circles on the inclined plane. We are inclined to call the trajector "above" the landmark, but not if the inclined plane is absent (though all else remains the same).

3. In the real world, a child often is told a correct label for a spatial relationship, but is not explicitly told which spatial terms *do not* apply ("The book is *above* the box, Johnny; it's not to the left of it, or to the right, or below, or..."). This wouldn't be a problem if the child could simply assume the mutual exclusivity of spatial terms; i.e. if he could conclude from the fact that the book is *on* the box that it isn't related to the box in any other spatial way. The problem is that the book is also *outside* the box; and in general, spatial-relation terms are not mutually exclusive. On the other hand, a system that receives *no* negative evidence will not learn.

4. Some spatial-relation terms exhibit "spatial invariance" across the input bitmap images. That is, across all spatial shifts of part of the image (e.g. the trajector), within some range perhaps, the same spatial-relation term is used; it is "invariant" to those spatial shifts in the input. So, e.g., say the trajector is *above* the landmark; then a little nudge of the trajector to the left or right will still leave it above the landmark. This isn't a problem *per se*; the problem is that Regier wanted to build a network which could generalize from some examples of (e.g.) *above* to all examples, i.e. could learn the spatial invariance without having seen every example of it. PDP networks cannot perform this task.

Regier's model has two main pieces: (1) a directional feature detector and (2) a non-directional one....

Regier's is a structured connectionist model, and it learns by backprop. We say "structured" because many of the feature detectors were simply built in rather than learned. This amounts to the assumption either that we are born with such detectors, or that we've been trained up on them prior to the acquisition of spatial-relation words. So, for instance, Regier built in directional-feature detectors like proximal orientation, center-of-mass orientation (relational features); major/minor axes of landmark, upright vertical (reference features). There are also feature detectors for topological relations. This was in effect Regier's solution to (4) above: he transformed the input space from bitmaps to features, in the processes transforming the spatial invariances into other features—features that a PDP network could generalize over.

The system is presented with examples of various spatial-relations images, and simultaneously a word which corresponds to it. This gives rise to the problem of encoding negative evidence, (3) above.

Here's Regier's solution: down-weight the error on (implicit) negatives; that is, make them count less than positive evidence. In mathematical terms, this means that the error term is changed. Recall:

$$E = \frac{1}{2}\sum(t-o)^2. \tag{2}$$

This becomes

$$E = \frac{1}{2}\sum((t-o)\beta)^2, \tag{3}$$

where β is a real number on the interval $(0,1)$. This worked.

Regier also enhanced his model to work on "movies" in addition to static images. This required an extra hidden layer....

14 Class #14 (3/2/06)

14.1 Regier model: denouement

Finally, a word about what Regier's model *can't* do.

1. It doesn't scale up very well;
2. it can't do any kind of grammar production; it's not provably the unique solution to this problem;
3. it can't perform inference ("if x is to the left of y then y is to the right of x");
4. it can't learn abstract (as opposed to concrete) meanings of spatial term;
5. it's representationally opaque.

On the other hand, it demonstrates the utility of inductive biases (without which the program could not have learned); it makes use of topographic maps; it learned in spite of the lack of explicit negatives, as with real language learning.

14.2 Force dynamics

We looked briefly at Talmy's schematic representation of force dynamics. The general idea is that we can think of, or model, causal sentences—i.e. some subset of force-dynamic sentences—in terms of three primitives:

1. agonist/antagonist
2. initial tendency
3. final result

We looked at some sentences and their schematic representation, and then at some other sentences which had force-dynamic representations in various semantic fields: physical, psychological, socio-psychological etc.

Image schemata and force dynamics are *closed-class* terms, meaning "a word class to which no new items can normally be added, and that usually contains a relatively small number of items. Typical closed classes found in many languages are adpositions (prepositions and postpositions), determiners, conjunctions, and pronouns" (from the Wikipedia). Can we do similar analyses for open-class terms? Recall, an "open word class, in linguistics, is a word class that accepts the addition of new items, through such processes as compounding, derivation, coining, borrowing, etc. Typical open word classes are nouns, verbs and adjectives" (from the Wikipedia.) Our answer, as you may guess, is "yes," but the basis for these will be *motor* schemata.

Motor schemata are used to execute behavior. Our claim is that motor routines are schematized, meaning that there are generalized motor schemata which are used for a variety of tasks. This is critical for our stronger claim that these same circuits are used to process language. We will revisit this theme shortly.

14.3 Motor areas in the brain

Motor area F5 contains neurons which fire for, e.g., generalized grasping; i.e., whether the grasping is done with the hand or mouth or whatever, these circuits fire (but presumably not for, say, pulling). On the other hand, F5 also contains neurons which control distinct phases of grasping (opening fingers, closing fingers, etc.) Finally, mirror neurons in F5 also exhibit the type of "universality," or generality, that we hypothesize underlies language processing: they respond when the monkey performs goal-related hand actions as well as when it observes another individual performing a similar action. What does this say? It says that the mirror neurons encode a general notion of *purposeful action* (since they don't fire when there's no object) and of an *agent* (since they fire regardless of who performs the action).

The F4-VIP (ventral-interparietal) region in the motor cortex appears to represent stimuli in peri-personal coordinates. That is, sensory information (e.g. visual and tactile) have representations in the premotor area which appear to be for the purpose of performing actions. So, for example, certain neurons are responsive to both visual and tactile input, but only from *the same body region*; the neurons are insensitive to visual input from anywhere outside the patch of skin from which they receives tactile input. The boundaries of peri-personal space are set by the workspace of the effectors of the human body. This appears to be a direct example of an integrated, multi-modal neural representation for the purpose of performing an action.

There are also cells in the somato-sensory areas which, similarly, respond to activity in specific body-space regions—independent of modality (well, it works at least for sound and sight).

Finally, the PMv (F5ab)-AIP circuit contains neurons which fire in relation to affordances; so, for example, they encode size of objects.

Somatotopy for effector actions, in both observation and execution scenarios, has also been observed in fMRI studies in humans. (All of the above information is from monkeys).

Our conclusion: these motor circuits are used not only in exectution but in the *representation* of actions. What does that mean? It means that when we think and talk about actions—and, as we shall see, about abstract domains which draw on these actions through metaphor—we use these circuits.

15 Class #15 (3/7/06)

Today we celebrated what we've learned by taking a midterm.

16 Class #16 (3/9/06)

We talked about four main components of motor control schemata:

1. CPGs
2. reflexes
3. command signals from higher motor sensors (BG, motor cortex, etc.)
4. la la la

And about some requirements for motor control computation:

1. coordinated, distributed, parameterized;
2. active (i.e. it's a *circuit*, not a static structure);
3. able to model *concurrency* and *interrupts*;

4. should model *hierarchical control*.

Our model which captures these features is called the *execution schema*, or x-schema for short.

16.1 X-Schemata

Execution-schemata are a computational framework for modelling actions. In fact, an X-schema is really just a stochastic Petri net, which is a long-standing mathematical representation in computer science. If you're not familiar with them, you should read the Wikipedia entry, and take a look at some of the documents at Petri Net World. The short story is that Petri nets are "a mathematical representation of discrete distributed systems." In some sense they generalize finite state machines (though the relationship between them is actually somewhat more complicated).

The major difference between standard Petri nets and x-schemata is that the latter may have more complicated firing functions: In addition to conjunctive firing (i.e., the resource requirements of all the inputs are satisfied), there is stochastic firing, where the probability of a transition is given by the logistic function of (say) the sum of the inputs. Alternatively, the stochastic function may be chosen from the exponential family.

Obviously x-schemata should be able to (1) support the four components of motor control, and (2) satisfy the four requirements just listed.

17 Class #17 (3/19/2006)

Dr. Feldman taught this class.

One way to think about language learning and processing is in terms of a "constrained best fit" paradigm. The idea is that we try to find the best fit between our perceptions (e.g. the labelling of an action with a word) and an internal model (e.g. of grammatical rules, or of the use of a word, etc.). The model, of course, can be updated or amended. The mechanism by which the best fit is calculated is usually probabilitic. Finally, we stressed that the best-fit calculation may weight the outcomes by their relevance to the creature's (our) plans and goals; so for instance it may be "liberal" in labelling certain animals as wolves, since its survival is better ensured by favoring false positives over false negatives.

David Bailey built a model in this paradigm which was to learn labels for actions, and also to be able to carry them out when so commanded. It consisted of three different computational layers: x-schemata, "feature structures," and a set of probabilistic computations for updating the model in light of new evidence (labels).

Hypothesis: "Languages only encode those parameters of which we're consciously aware." That is, we can only talk about things we're consciously aware of. This probably seems sort of trivial [it does to me, too, frankly], but the idea is that our words encode parameters into actions, and these parameters are the things we can consciously control. As we become aware of some new parameter ("Ah, I can modulate the angle at which I throw this!"), we learn new words for it, or some new way of encoding it in our language.

In Bailey's model, these parameters are encoded in an intermediate layer of "feature structures," each of which lists both the parameter (e.g. hand position) and its value (e.g. index-finger extended, or flat palm, or etc.) for all the parameters associated with that word. The feature structures are hooked up to the x-schemata. The idea is that the x-schemata are already in place—they are our sensorimotor circuits, after all—but they are *parameterized*, i.e. they execute differently depending on the value of parameters. These are exactly the parameters which are encoded in our language, which we're consciously aware of, and which are represented in the feature-structure layer. In fact, which x-schema gets selected is itself represented as a parameter. Here's a list of (some of) the parameters in the Bailey model, and their possible values:

- which x-schema?
- hand posture (grasp, palm, index finger, etc.)
- direction (toward, away, up, down, etc.)
- elbow joint motion (flex, extend, fixed)
- force (low, medium, high)
- aspect (whether the x-schema repeats)
- object size (small med large)
- depressability (etc.; this is one of a few object properties)

17.1 Intermezzo: fitting models to data

There is in general a trade-off between simple models that fit the data rather poorly but can generalize to new data well; and complicated models that fit the data very well but generalize to new data rather poorly. This is in fact something of a general problem, and we'll discuss one fairly simple approach to solving it. The main idea comes from a concept in statistics called Bayes' Rule. We need to take a brief tour through some very basic probability facts in order to explain it.

The conditional probability of A given B, written $P(A|B)$, is the probability of event A occurring, given that event B occurred. Of course, sometimes events A and B are unrelated, so that $P(A|B)$ is simply $P(A)$. In this (special) case, the events A and B are said to be *independent*. In general, though, this is not true. Now, conditional probability is defined in terms of unconditional probabilities:

$$P(A|B) := \frac{P(A,B)}{P(B)}. \tag{4}$$

Here $P(A,B)$ is the probability of both A and B occurring. It is highly recommended that the reader unfamiliar with probability convince himself that this formal definition accords with the intuitive definition just given. The reader should also note that in the case where A and B are independent (written $A \perp\!\!\!\perp B$), which is formally defined as

$$A \perp\!\!\!\perp B \Leftrightarrow P(A, B) = P(A)P(B), \tag{5}$$

then in fact $P(A|B) = P(A)$. Independence is symmetrical, so $A \perp\!\!\!\perp B \Leftrightarrow B \perp\!\!\!\perp A$.

Now if we use this definition of conditional probability (twice: once for $P(A|B)$ and once for $P(B|A)$) and rearrange our terms a bit, we get Bayes' Rule:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \tag{6}$$

What's so special about Bayes' Rule? It might seem rather trivial, since after all it's just a rearrangement of the definition of conditional probability. However, it is often the case that we know the probability of an event, A, given that another event, B, happened, but have no *a priori* estimate for the opposite dependence; and Bayes' Rule provides a method for computing exactly that, assuming that we also know the so-called prior probabilities of A and B, namely $P(A)$ and $P(B)$. This often happens when dealing with causal situations (though importantly, Bayes' theorem is in no way restricted to these cases).

So, e.g., I may know that a hard rain is likely to soak my kitchen floor through the leak in the window sealing; maybe this happens four out of every ten "hard" rains, so I estimate that the probability of the kitchen floor getting soaked (S) given that it rained the night before (R), $P(S|R)$, is 0.4. It seems reasonable to suppose I also know that in the course of the year my kitchen floor get soaked about ten times, so we estimate that the prior probability of it being getting wet, $P(S)$, for one reason or another (maybe the faucets explode every now and then, too), is $10/365 \approx 0.03$. Finally, I check the meteorological data and see that there are 15 hard rains every year, so I take the prior probability of hard rains, $P(R)$, to be $15/365 \approx 0.04$. Now, waking up and finding my floor wet, I have no idea what the probability that this was caused by a hard rain is—but I can just plug my numbers in Bayes' Rule, and get exactly such an estimate, in this case just over 0.5, which says that a little more than half the times my floor is wet it was probably caused by the rain coming through my poorly sealed windows.

How does this relate to fitting models to data? Well, in general we might think of the task of model fitting as the task of choosing the model M that's most probable, given the data, i.e., $\operatorname{argmax}_M P(M|D)$. The problem is that we don't know *a priori* how to compute $P(M|D)$—and the solution is to use Bayes' Rule:

$$\begin{aligned} \underset{M}{\operatorname{argmax}}\{P(M|D)\} &= \underset{M}{\operatorname{argmax}}\left\{\frac{P(D|M)P(M)}{P(D)}\right\} \\ &= \underset{M}{\operatorname{argmax}}\{P(D|M)P(M)\}, \end{aligned} \tag{7}$$

where the last equality follows from the fact that the denominator $P(D)$ is not a function of the model, M, and hence doesn't affect the argmax. So all we need is some way to estimate the prior probability of the model, $P(M)$, and the so-called likelihood, $P(D|M)$. In the model fitting task we can think of $P(M)$ as being proportional to the simplicity of the model; i.e. the more complicated the model, the less likely it is. This term weights the model choice in favor of simple models. The likelihood, on the other hand, weights the model choice in favor of the one that fits the data best, or in other words that makes the data most likely. Lastly, $P(M|D)$ is known as the posterior probability; hence the quantity on the left-hand side in Eq. (7) is known as the *maximum a posteriori probability*, or *MAP*.

Finally, a common technique when performing these kinds of analyses is to take logarithms before applying the argmax function; otherwise we end up multiplying very small numbers by each other and leave our computations susceptible to round-off errors. So instead of writing our maximization task as in (7), we write

$$\log[P(M|D)] = \log[P(D|M)] + \log[P(M)] - \log[P(D)], \tag{8}$$

so that when we take the argmax—and again note that the last term doesn't affect the argmax—we get

$$\underset{M}{\text{argmax}}\{\log[P(M|D)]\} = \underset{M}{\text{argmax}}\{\log[P(D|M)] + \log[P(M)]\}. \tag{9}$$

Note that this equation can just as easily be written in terms of minimizing a cost, if it pleases us (and as you will see in your homework assigment); we just need to invert everything back at Bayes' Rule, Eq. (6), and then take an *argmin* over the inverse of the posterior probability.

18 Class #18 (3/16/06)

Dr. Feldman taught this class, too.

18.1 The Bailey model resumed

We looked at an example of two different feature structures, each of which encoded a sense of the word "push": namely, pushing something and moving it, and pushing against (e.g.) a wall (and not moving it). You should look back at how the values for the various parameters differed across these two sense of word "push."

Here's the child-developmental basis for the Bailey model. A child learns action words (it is hypothesized) when he performs an action, and hears it labeled by his mother. But the same word can have different senses; i.e., the same word can be used to label (fairly) different actions. On the other hand, mild variance in the parameters should *not* indicate a new word sense. So, for example, the child may push something slightly harder than usual and still have it labeled as a push, and he should not think of this as a new sense of the word.

The trick is: when hearing a word, choose in the "best way" whether to create a new sense for that word, or instead to broaden the range of a parameter (or parameters) that count as an example of the original word sense. It's particularly tricky because in fact we can get cases where one parameter (or several) take on several radically different values, but we'd like the program not to make several different word senses, but to just throw out that parameter as meaningful to the word at all.

This way of incorporating new examples into an existing model in order to best account for the data is called "model merging," and Bailey's method of performing the trick just mentioned was to find the maximum a posteriori probability (MAP) of the model given the data, i.e. the technique described at the end of the last lecture. It worked pretty well (and for the details of its success, see the lecture notes).

By the way, the Bailey model wasn't actually randomly performing actions and getting them labeled, like a child; rather, an action and its label were input as a feature structure and, well, a label. The feature structure was meant to represent the action that the child has just performed, and the label as what the child heard from a competent language speaker.

Here are some of Bailey's conclusions:

- high-level control is most relevant
- universal constraints render learning tractable
- features link action to reasoning
- the Bayesian formalism works

Finally, we talked about the details of how the Bailey model uses the Bayesian formalism to perform model merging. In fact, Bailey used it twice, once to learn the best model for the given action and its label, (i.e., create a new verb sense or not); and once after training to find the verb that best describes a given experience (and given the trained-up model, of course). The first of these is written mathematically as: ..... ; and the second as: .....

19 Class #19 (3/21/06)

In this lecture we talked about the structure of events as they are represented in language.

"Many inferences about actions derive from what we know about executing them." This is one of our (major) hypotheses about the neural basis of language and thought. So at least some of the neural circuits which are involved in executing an action (or actions) are also used in reasoning about that action (or actions). As we have seen, we model these circuits with a form of modified Petri nets known as x-schemata. X-schemata give us the following necessary features for executing actions, which also happen to be very useful for reasoning and inference drawing:

- parameterization
- dynamic binding
- hierarchical control and durative transitions
- stochasticity and inhibition

19.1 Aspect

The ways in which lexical and grammatical devices encode the structure of events in a language are collectively called *aspect*. In English, tense and aspect are often conflated, but the distinction is this: tense encodes the absolute position in time of an action (something like placing it on a time line), whereas aspect encodes how the event *is to be viewed* with respect to time or other events. In particular, apsect distinguishes among the following:

- viewpoints: *is walking* vs. *walk*; the action is in the present tense in both cases, but should we view it as ongoing (the former) or habitual? There are other viewpoints as well.
- phases of events: *starting to walk*, *finished walking*, etc. This is self-explanatory.
- inherent aspect: *cough*, *rub*, and many other verbs exhibit an inherently iterative aspect. Others, like *sneeze*, are inherently punctate—though this can be changed under the influence of modifying phrases. (So, e.g., in the sentence "She sneezed for five straight minutes," the aspect of "sneeze" has been shifted to inherently iterative.
- composition with...

Now, linguists don't all agree on what counts as aspectual in English, or even what constitutes aspect in general. So take this as one particular account, though details may vary in other sources.

Here are some examples of aspect:

- **Perfective**. A verb in the perfective aspect includes its beginning, middle, and end; i.e. it can be thought of as a point on a time line. For example,

  (1) John entered the room.

- **Imperfective**. A verb in the imperfective aspect, on the other hand, is viewed as "smeared out" over the time line, and its beginning and ending are not in view:

  (2) John was singing.

  Imperfective aspect is sometimes thought of as subdividing further into habitual and progressive aspects.

  - **Habitual**. Habitual aspect indicates that the action took place frequently over some period of time, or that it took place over some protracted period of time. In English this really only occurs in the past tense, and employs the auxiliaries *used to*:

    (3) John used to beat his children.

    (4) John used to think that the earth revolved around the sun.

  - **Progressive**. On the other hand, the progressive aspect indicates that an action is ongoing, as in

    (5) John was playing cards.

- **Perfect**. The perfect aspect (not to be confused with perfective!) indicates the temporal relation of the action to some other action or event. Consider:

  (6) It was clear that John had seen through me.

  The perfect aspect of *had seen* indicates the temporal relation of John seeing through the speaker with respect to the time that the sentence was spoken (or thought or whatever), namely that the former took place before the latter. Whereas if the aspect were neutral as in

  (7) It was clear that John saw through me.

  the seeing through may have taken place before the sentence was uttered, or it might be taking place as the sentence is uttered. We don't know. The past tense is used in these two sentences, but the contrast can be effected as well in the present tense:

  (8) It is clear that John has seen through me.

  (9) It is clear that John sees through me.

- **Prospective.** The prospective aspect indicates that an action is about to happen. In English, this is effected with the use of an auxiliary phrase like "to be about to," or "to be on the point of," as in

  (10) I was on the point of striking him when he finally stopped misbehaving.

- **Inceptive.** Some languages have markers on verbs to indicate their beginning or inception. In English we use the auxiliary "started":

  (11) She started knitting.

- **Iterative.** We mention one more aspect mark, though again linguists make other distinctions as well. The iterative aspect is indicated in English by various constructions particular to various verbs.

  (12) They ran around the track twice.

  (13) She sneezed for a almost a minute.

  (14) He refused to stop hitting me.

  Notice that without the prepositional phrases in (12) and (13), the aspect would be imperfective and perfective, respectively. Actually, in *They ran*, the imperfective aspect is *lexicalized*, i.e. it is built into the meaning of the word *ran*, which is why no auxiliaries are necessary. In (14), the word *stop* indicates that the action is ongoing, whereas the lexical (as opposed to grammatical) aspect of *to hit* is normally perfective; the interaction sets up the iterative aspect.

Much of the previous discussion, including some examples, was drawn from this excellent, very short introduction to aspect, which is very highly recommended reading.

19.2 X-schemata and aspect

Now actions seems to exhibit a very general structure: they have beginnings and ends and ongoing periods, as well as states of suspension, having been canceled, having been finished, and being ready to begin; and they can be iterated, resumed, canceled, and interrupted. We looked at a general x-schema, called the controller x-schema since it controls other actions, which has just these properties. And now we note that the aspect of a sentence seems to correspond to which state in the controller x-schema the action is, or which transition is taking place. That is, the distinctions that linguistic aspect makes correspond to stages in the execution of the the controller x-schema. The reader is strongly encouraged to review the slides on the controller x-schema to see exactly which aspectual markers are mapped to which execution states or transitions of the controller x-schema.

Our hypothesis: in order to understand the aspect of a sentence, we execute the controller x-schema up to the appropriate stage—though we execute it while "disconnected" from any particular action. We call this execution of a (sensori-)motor circuit without execution of the action *simulation*.

There are classical characterizations of aspect, but in class it was proposed that the generic control features from the controller x-schema actually provide a more useful framework for thinking about aspect.

Reichenbach proposed a way of parsing tense and aspect according to the relative placement on a time line of *event time (E)*, *speaker time (S)*, and *relative time (R)*. We showed a way of combining this system with x-schemata to represent both tense and aspect.

20 Class #20 (3/23/06)

The subject of today's class is frames, and the FrameNet project. The lecture was given by FrameNet's Michael Ellsworth.

A frame is a conceptual unit akin to schemata, though they are culturally specific. Words evoke frames, and with it the concomitant frame elements. Our classic example is the commercial transaction frame, which is evoked by (e.g.) the word "buy," and has frame elements like "buyer," "seller," "good," etc. The FrameNet project basically goes through the language (English and others) and picks all the different frames associated with each. They also annotate sentences, tagging each word with the frame that is evoked by each word in that context. The project is ongoing, but all current results are available from their website.

21 Class #21 (4/4/06)

21.1 Metaphors as maps

Today's lecture was on metaphor. We went through the lecture fairly fast, and there's lots of information that didn't make it into these notes. You shoud consult the slides.

The main claim of which we want to convince you, and which the rest of the class was meant to support, is that our understanding of abstract notions nevertheless employs the same sensorimotor circuits we've been talking about all along. This is accomplished by way of conceptual metaphors, embodied in brain circuits, which connect abstract representations to a concrete substrate. We call these "metaphoric maps."

What's so nice about this setup? Well, it allows us to reason about things in the abstract domain without additional apparatus. Specifically, it gives us the tools to draw inferences and to make correlations.

Formally, metaphors are mappings from the concrete (source domain) to the abstract (target domain). There is a schema (we claim) for the abstract notion (e.g. emotional state) and a schema for the concrete notion (continuing the example, say degree of verticality). Then there is a circuit which combines the two.

How do these maps get set up? Well there's a set of so-called primary metaphors that we claim are established by the concomitance of the abstract and concrete experiences. This claim is called the "conflation hypothesis." Examples:

- *Affection is warmth.* When we experience (a display of) affection, we often also experience warmth, as when being hugged. This gives rise to metaphors like, "She gave me a cold look," and "I received a warm reception."

- *Important is big.* In the childhood environment, big things are often the important ones—like, e.g., one's parents. So we get metaphors like, "Tomorrow is a big day."

- *Categories are containers.* Things that go together tend to be in the same bounded region. This gives rise to metaphors like "The whale doesn't really *fit with* fish, because it gives birth to live young...."

Here's a more specific description (Chris Johnson's) of the acquisition of these primary metaphors:

1. The source domain word is used in the source domain.

2. Certain constructions acquire double meanings.

3. Constructions specific to the target domain appear.

One (nearly?) ubiquitous metaphor is the so-called event-structure metaphor. It consists of many mappings:

- states are locations

- changes are movements
- causes are forces
- causation is forced movement
- actions are self-propelled movements
- purposes are destinations
- means are paths
- difficulties are impediments to motion
- external events are large, moving objects
- long-term, purposeful activities are journeys

The slides have examples of each of these. What's neat is that there are all sorts of entailments in the abstract domain which are mapped from entailments in the concrete domain. So, e.g,

- making progress is forward movements;
- amount of progress is distance moved;
- undoing a process is going backward;

and many others.

More metaphors:

- *Ideas are possessions.* "She stole my idea."
- *Ideas are resources.* "He ran out of ideas"; "Let's pool our ideas."
- *Ideas are external entites.* "A new idea took hold of me"; "I shied away from those memories."

21.2 Boroditsky's psycholinguistic experiments

We'd like to convince you that the way you reason about target (abstract) domains really does draw on or access the source (concrete) domains. So we look at some experiments performed by Lera Boroditsky (Stanford).

There are two time metaphors that get used regularly in English. In one, the ego (self) moves through time, which is a stationary frame of reference. In the other, the ego is stationary but time is moving toward the ego.

Now think about the sentence, "Next Wednesday's meeting has been moved forward two days." When do you think the meeting is? You may have clear intuitions, but it turns out that people give two different answers, Monday or Friday, depending on which of the two time metaphors they're using. That is, if they're thinking of the ego-moving metaphor, then "forward" means in the direction of the motion of the ego, which is Friday; whereas in the case of the stationary ego, "forward" means the direction in which time is moving, which is Monday.

Now here's the neat part. It turns out that you can prime subjects to use one or the other metaphor by having them either (say) walking (ego-moving) or pulling an object toward themselves (ego-stationary). In about 70% of the former cases, the subjects think the meeting is on Friday, and in about 70% of the latter they think it is on Monday. So it looks like we really do use our sensorimotor circuits in reasoning about time, since priming the former affects the latter.

Moreover, it turns out that the priming doesn't work in the other direction. If you prime a subject with one time metaphor, it has no effect on his perception of spatial relations.

Interestingly, vertical spatial primes also increase reaction times for time-based questions ("Is March earlier than April?") for Mandarin speakers, who have a verticality-based time metaphor; and for English speakers who've been trained with a vertical time metaphor; but not for normal English speakers. This isn't conclusive, but it's very powerful evidence for the use of concrete domain circuits, specifically sensorimotor circuits, in processing abstract domain sentences.

22 Class #22 (4/6/06)

22.1 A computational model for metaphor?

Here's one of our major hypotheses—maybe *the* major hypothesis—about language: We understand utterances by mentally simulating their content. Now, "simulating" is a term of art here, but it basically means that we exectute the circuits involved in carrying out (e.g) the action we hear described. So if I want to understand the sentence, "John walked into the cafe," I use some of the same circuits I myself would use to walk. This is a strong thesis. But our contention is even stronger: we claim that even in understanding sentences that don't literally describe anything to do with the sensorimotor (or affective) systems, we nevertheless use sensorimotor (and affective) systems.

This is a very strong claim indeed, and last lecture we presented some of the psychological evidence for it. Now we'll describe a computational model for it, one which draws inferences in abstract domains via mappings to concrete (sensorimotor) domains, where correlations and inference-drawing mechanisms are already in place. To re-stress the central idea: The reason these mechanisms are already in place in the model of the concrete domain is that this model is supposed to represent the sensorimotor circuits that we are either born with or which quickly develop during the early stages of our interaction with the environment.

How do these computational circuits "simulate" sentences? Well the linguistic structure of the sentence (and the context) provides *parameters* to the computational circuit. This works just as we described for the Bailey model. (You should go back and revisit that now if you don't remember.)

22.2 Bayes Nets

Classically, inference-drawing was thought of as deductions in formal logic. You'll notice that this fits in with the so-called classical notions of categories and of concepts that we derided in Class #11. Not surprisingly, then, we don't think this is the right way to approach inference. We (humans) seem to draw inferences that are *weighted*; i.e. we believe some things more or less strongly, with greater or less probability. So we'd like a model of inference that quantitatively combines evidence in context to draw the most likely conclusions. And given the psychological data we talked about in Class #5, we think the constraints which get combined can come from anywhere, i.e. they're not simply bottom-up, or simply top-down.

This approach also maps nicely onto our neurological architecture: Imagine that some local collection of neurons (a "node") represents an event or a fact. And then say that this node is connected to other nodes which represent facts whose truth or falsity is dependent, probabilistically, on the truth or falsity of our first fact. And now, finally, imagine that how "active" this node is, that is how much the underlying neurons are firing (assuming a simple rate coding), is proportional to one's belief in the truth of this fact. This is congruent with our understanding of neural architecture—though it is a giant leap of logic to the notion that groups of neurons represent "facts" or "events," since we have no evidence for this; and to the notion that degree of activation corresponds to degree of belief in its truth.

Well fortunately there is a computational framework which corresponds precisely to the one just (loosely) described in neural terms. The models are called Bayes Nets, after the same Thomas Bayes whose work in probability theory we saw earlier. Rather than re-present Bayes nets here, you should read the Wikipedia entry: `http://en.wikipedia.org/wiki/Bayes_net`; and also Kevin Murphy's tutorial, which has information on dynamic Bayes nets as well: `http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html`.

22.3 KARMA: A computational model for processing metaphoric sentences

The input and output of the system are feature structures. (Recall these from previous lectures, especially the Bailey model.) The source (concrete) domain is modelled with x-schemata, which provide for both representation and inference. The abstract domain is reprented by dynamic Bayes nets. These two domains are connected by three different kinds of metaphoric *maps*: parameter maps (pmaps), ontological maps (omaps), and schema maps (smaps).

source f=struct target bayes nets

pmaps parameter maps omaps ontological maps (roles) smaps schema maps

23 Class #23 (4/11/06)

Prof. Lokendra Shastri gave today's lecture. Here're the main questions: How does the a vastly parallel system of neurons encode a large body of episodic, semantic, etc. information? How does it perform inferences on sentences on the order of hundreds of milliseconds? How can it process 150-400 words per minute? Dr. Shastri referred to these types of reasoning as *reflexive reasoning*. The model which attempts to solve just these problems is called SHRUTI. [I take it this is an acronym, though I have no idea what it stands for.]

Imagine you here the sentence "John gave Mary the book." Now we've said (while waving our hands) that entities are represented by small populations of neurons. So hearing this sentence causes the "John neurons," the "Mary neurons," and the "book neurons" to fire. Now the question arises: How do we know that John is the giver, and Mary the receiver, and the book the entity that was transferred? I.e., how does the system *dynamically bind* entities to roles? Now if we hear the sentence "She threw it away," how do we know that "it" refers to the book? I.e. how does the system propagate dynamic binding?

—computing coherent explanations and predictions—

one-shot learning of events and situations gradual/incremental learning of relations, schemata, and causal structures (SMRITI)

Here's the proposed solution: Say that each entity, category, and relation is represented by a *focal cluster* of neurons. Now there is, e.g., a focal cluster which represents *giving*. Within this cluster, there are neurons which represent *giver*, *receiver*, and *given* (and frankly, maybe more things). So: to represent the sentence which began our discussion, the *given*,*giver*, and *receiver* neurons in the *fall* focal cluster fire, as do the neurons representing *book*, *John*, and *Mary*, and perhaps they even fire at the same frequency. But—and this is the key—the corresponding neural populations are phase-locked together; e.g. the "book neurons" fire at in phase with the *given* neurons, and with no others.

Building on this solution, we can imagine permanent (one-way) links which build up between, e.g., the representations of "John" and "man." This allows us to *generalize* from specific examples. We should also note that focal clusters become active when we perceive an event, remember one, understand a sentence about one, or experience one.

The reader may be wondering if we have any neurobiological or pscyhological evidence for this mechanism. The answers (I think) are no and yes, respectively. The psychological evidence goes as follows: In EEG readings, activity moves between various frequency bands, and in particular we see waves of synchronous in the so-called gamma band, 25-60 Hz. If we take the low end, 25 Hz, we get a period of 40 ms. How much information can we pack in here using phase-locking? Well, say we can tolerate a jitter of about 3 ms; then we can only differentiate between waves of about seven different phases. (If this picture isn't clear, consult the slides. The idea is that, if we tried to "squeeze in" a wave of a different phase, it would be too close to its neighbors and so might be mistaken for them, and our binding would fail.) What's really neat about this result is that seven seems to be about the number of bindings that human beings can keep track of at one time; i.e. we couldn't keep track of bindings in a sentence which introduced more than seven of them.

Here are the main claims about Shruti:

- An episode of reflexive reasoning is a transient propagation of rhythmic activity.
- Each entity involved in this reasoning episode is a phase in this rhythmic activity.
- Bindings are synchronous firings of cell clusters.
- Rules are interconnectioins between cell-clusters that support context-sensitive propagatioin of activity.
- Unification corresponds to merging of phases.
- a stable inference (explanation, answer) corrsponds to reverberatory activity around a closed loop.