Inference and Probabilistic Modeling: Week 2 (William)

***Probabilistic models of cognition: where next?**: Chater, Tenenbaum and Yuille*

Premise: Advances in probability theory and statistics allows scientists to reconceptualize cognitively-relevant information processing.

They cover the challenges in three domains: Representation, processing and learning

Given: Cognitive systems are complex, require a rich, composition system of representation e.g. hierarchical structure of motor system; multiple layers of complexity in visual images; and phonological, syntactic and semantic regularities in language.

Solution: Probabilistic models are increasingly able to encompass such richness. Reduction of systems to simplistic representations is no longer necessary (i.e. gradient systems reduced to binaries).

Task: The challenge is to find a way to treat atomic states as structured (a state embedded in a system e.g earthquake yes/no becomes earthquake, duration, intensity, location).

Processing:

People have trouble with explicit probabilistic processing (see work by Kahneman, Tvesky, Giggerenzer ec.). Does this mean that our cognitive systems cannot be probabilistic? Computational research shows that Bayesian networks map well to distributed, parallel processors (which is what we surmise the connectionist, neural network may be).

- *Assume Bayesian probability can be implemented, the question then becomes, scope. What would the reach of such probabilistic architecture be? Can it be recruited? What kind of processing would occur?*
- *This has implications for how we view reason- is it Bayesian (i.e. rational and able to apply logic)? How would a network built on causal dependencies accommodate suppositions, counterfactuals and other preferential or non-true modes of reasoning?*
- *Possibilities: Yuille and Kersten propose perception as analysis by synthesis: probabilistic updating is some kind of MCMC process.*
- *Question: how would MCMC or other P model map onto neural hardware as we know it?*

Learning:

Inferential learning implies that we learn structure (and principles and benchmarks) from data. We infer relationships between behavior, environment and reward.

We have made progress in inference from concrete objects. Greater hurdles remain in inferring representational structures: the space is enormous and discontinuous; calculation seems computationally prohibitive; structures may be contained in highly abstract ways.

But: do we have enough data to really learn structures at all? Are there innate constraints that make learning possible (must have some sort of discrimination function that lets us encode learnable, ignore unlearnable).

Do models fit neuroscientific and behavioral data?

Comments:

This sets up exploration of Bayesian models as they relate to syntactic tasks: language processing and acquisition.

***Probabilistic models of language processing and acquisition**: Chater and Manning*

Inference and Probabilistic Modeling: Week 2 (William)

State of linguistics: After (or beginning with and following) Chomsky, focus on symbolic represetations: trees, matrices, logical representations. No room for or acceptance of probabilistic processing.

> *Point of Emphasis: Probabilistic approach to cognitive systems assumes that the structure of information processing itself is probabilistic. This is markedly different from Yang's characterization of probability in linguistics as simple frequency distributions within categories/boundaries set by absolute linguistic rules (e.g. innate grammaticality test). Prob/IP means we infer the structure whereas classical linguistic structure means the structure is given (as UG).*

*Their Argument: language is represented by a probabilistic model and that language processing involves **generating** or **interpreting** using this model, and that language acquisition involves learning probabilistic models. They focus on parsing and learning grammar.*

> *Thus, probabilistic systems complement linguistics but they also offer a way to account for the cognitive aspects underlying language (as information process). C&M wish to explore this latter aspect.*

Overview of the Probabilistic Model = How it works

Chomskian grammar: system of rules that specifies all and only allowable sentences.

Information-processing: Disambiguate data stream (discern words, syntax, grammar)

> *Parsing: to infer underlying linguistic structure (tree) from string (or stream) of words S where t is an element of T (this branch is an element of the master Tree) given that s is one (of countless) instances of all possible Strings.*

Problem: how to parse correctly? C&M focus on syntax rather than phonology (group words rather than sounds. For sounds, see Tenenbaum, Griffiths and Kemp).

Parsing improves in speed and accuracy if the bracketing system is known (operates automatically):

> *[the [old [man]]]= I read this as three discrete words with old modifying man (noun)*

> *[[the old] man]= I read this as the old (elderly) man (verb). Here old refers to a concept and man requires inferring secondary meaning*

That is, knowing how to bracket prunes the space to be processed. The question is how would a probabilistic framework account for the bracketing?

They extend Bayesian approach that seems to support research into vision, inference and learning to language:

"Information about the probability of generating different grammatical structures, and their associated word strings, can be used to finer grammatical structure from a string of words."

The parsing problem is now formally noted as:

> *$Pr_m(t|s)$ or the probability of a tree given a sentence s and given a probabilistic model Pr_m*

Question: How do we come up with the correct probabilistic model given that it could be anything?

- *C&M adapt idea from psycholinguists that we prefer first reading of a sentence due to minimal attachment (attach less to each word/drill down less), which offers a speed advantage.*
- *Probability of a tree is the product of the probabilities at each node; and hence other things being equal, fewer nodes imply higher probability (337)*

Inference and Probabilistic Modeling: Week 2 (William)

The girl saw the boy with the telescope:

- *girl saw: boy with telescope*
- *girl: saw with telescope: the boy*

Problem: This could lead to serious reduction: fewer nodes=higher probability thus always parse for simplicity. This problem can be investigated empirically through corpus such that a structure with fewer nodes but highly improbable rules would be dispreferred.

- *does dispreferred mean it would not likely attach as a prior belief/model in Bayesian logic due to low success rate?*
- *How do we estimate probabilities of each phrase in a stream? Corpus statistical analysis may reveal the probabilities (parsing frequencies of words that go together). BUT, how do we know these probabilities? Is it incremental learning? Geometric?*

In terms of representation, there may be a larger problem: How to deal with lexical information (meaning in the words, parsing of concepts to make sense semantically?).

Ideally, probabilistic information processing of syntax is ***context-free****. That is rules are rules and they are right simply, not on the basis of context (utterance or phrase is or is not allowed on the basis of rule application not on interpretation. This may be a legacy of linguistics, which uses grammatically judgments to discern symbolic, binary rules).*

BUT: change in words may invert tree structures or place constraints (replace TELESCOPE with BOOK)

Solution: Prob. models may incorporate head words as a lexicalized grammar: words beginning trees (load) likely candidates for subsequent groupings = X co-occurs with Y frequently and infrequently with Z.

Comment: We are no longer parsing individual words in a string but doing preferential bracketing (or perhaps a Bayesian bracketing wherein prior beliefs/model are updated based on current (new) information such that the next parse is more accurate than the last parse.

BUT: would this become a matter of trying to store too many correlations? What is more taxing-pruning space with auto-correlation or fixing after wrong auto-correlation?

Problem: Griffiths and Yuille note that calculating probability of joint distribution over multiple variables increases exponentially (i.e, phones in a stream or words in a sentence):

4 variables: 2^4-1=15 but 8 variables: 2^8-1=255 (computationally unlikely or impossible?)

One solution: G&Y: graphic representation using nodes reduces variables by their relationship and potential state (15 becomes 8). Each node is linked (maybe not nec. causally to another node)- MCMC shows how nodes may be related. But does this mean we need to also infer hierarchy (or at least conceptual potentials of each node) or is hierarchy built-into frequences of co-occurences of words?

Second Solution: Chater, Crocker and Pickering suggest that choosing most specific parse is best because it allows for fastest test of validity (correct on-line and in line rather than down the line).

- *But in the example "John realized his"*
- *Faster parse might be: phrase after HIS is open*
- *Correct parse (realize is an intransitive verb, phrase after HIS is constrained)= slower process but correct first try*

Experiment: Could we show reader such sentences and time latency responses to see how quickly they choose one reading versus the other? This might show how we parse.

No answer yet. C&M note that Prob. Modeling reframes questions in ways cogsci can investigate but it does not repudiate linguistic theory.

Larger issue of language learning

- *How can Prob. Modeling be **successful** in language acquisition and learning? That is, we parse correctly so that we communicate successfully (this would imply transparent and seamless processing)*

Prob. lang. processing "presupposes a probabilistic model of the language; and uses that model to infer, for example, how sentences should be parsed, or ambiguous words be interpreted. But how is a model, or for that matter a traditional non-probabilistic grammar acquired?

- *Chomsky: child has a hypothesis-space of candidate grammars; and must choose , on the basis of (primarily linguistic) experience one of these grammars.*
- *Bayesian: each candidate grammar is associated with a prior probability; and these probabilities will be MODIFIED by experience using Bayesian updating. Learner will presumably choose a language with high, perhaps highest posterior probability.*

Succeed by borrowing from speech and other research in neuroscience and cognitive science:

Yuille and Kersten's analysis by synthesis approach proposes that:

- *We infer low to high representation (from sound to word) by reverse-mapping from high to low (previously known parse predicts next parse because previous parses have probabilties attached to them. Not starting each time from zero. Assuming we learn, then parsing sounds or strings is not mutually exclusive but conditioned by Bayesian updating.)*
- *In their studies of speech and psychoacoustics, speech is also part of our cognitive system and we may hear internally the sound of words we are about to utter publicly. Phenomenology of 'inner voices' (hear before we say) may provide a **feedforward loop** where information is shared between producing an understanding speech.*

Chomsky's problem: how to account for learnability as cognitive process?

- *poverty of stimulus says a child cannot possibly be exposed to all the sounds, words, patterns to account for her ability to communicate in a syntactically correct manner or to generate new phrases (also grammatically correct).*
- *The rules are inside the head somewhere. The feeling of right/wrong in grammatical judments hints of a faculty (a phrase simply isor is not allowed. Perhaps there is a feeling that a sentence is plain wrong?).*
- *But how does Chomsky account for the learning of specific sounds as words? He makes the task one of mapping: we know all the rules and just need to know the mappings for the language we use (English, Chinese, Spanish etc.). Also, is it realistic detach building of semantic knowledge from grammar rules?*

Conversely, probabilistic modeling can account for learning sounds as words, sentences from words and structure from sentences and perhaps, semantics from meta-mappings for words and sentences in combination.

C&M: Moreover, learning need only succeed with high probability, more often than not, over time, learning improves as base gets bigger (geometric effect?).

This may be seen in the success of computational L. studies of corpora. Parsing trees from corpora has led to good approximations to syntactic categories and semantic classes "through clustering words based on linear distributional contexts." (p. 341)

- *distribution of a word that precedes and follows each token of a type*

- *this is a context-based argument and assumes that we discern hierarchical relation (nodes again). Parsing using brackets injects structure. Words are tokens of the order (in the tree structure) and types (what the word may be) as implied by the brackets.*

Linear order posits a dependency relationship between words: word co-occurrence implies structure but correlation is not enough. C&M: Two constraints:

1. *Dependencies more likely between adjacent rather than far apart words*
2. *More likely for a word to have fewer rather than more dependents*

BUT: How would we account for variation in conventions found in locative languages (Russian, Latin etc.) where word order varies due to conjugation of nouns and verbs?

Conclusion

Prob. models provide a way to map meaning representations. They offer a framework for building and evaluating theories of language acquisition and for concretely formulating questions concerning the poverty of the stimulus. They account for learning as a process and extend research using Bayesian logic in other area of cognition to language. Thus, they provide a path of inquiry rather than a dead-end in which grammar is accepted as innate. Probabilistic approach offers one line (and perhaps more fruitful) of attack.

Final thoughts

How do classical or computational linguists account for shadings of meanings, interpretation, from the vividness of good poetry to the grayness of poor prose? Why can there be dispute over what one (really) said or wrote? Is this a performance issue or something inherent in language? Would one approach do better than another in accounting for built-in ambiguity (think Postmodernism)?

Questions Inference and Bayesian Probabilistic Modeling (William)

General problem of communication and of distinguishing the following:

Sentence: What is said. What is meant. What is meant to be said.

Speaker, hearer observer: Each is attempting to parse sentence for these three things (and perhaps, also what is relevant or important = assign value to meanings)

I. Semantics:

Ray: From the discussion I can understand how applying probabilities with word-stream / tree pairs can aid in quickening the parsing task, but in order to use probabilities for disambiguation, I don't understand how this can work. To apply this, would we have to assign probabilities between word-streams and semantic interpretations, then? Because in linguistic theory, **not all semantic ambiguities are represented structurally**.

It also seems like, if it weren't bad already doing corpus linguistics, that coming up with probabilities for mapping word-streams to semantic interpretations will suffer from the sparse-data syndrome whereby there aren't enough interpretations available to pick a more common or less common one. In a sample set of 6 million sentences, for example, the British National Corpus, we hardly have enough semantic information reflected through syntactic distribution to differentiate categories of words robustly. How then, would we expect to coming up with this mapping between word-streams and semantic interpretations? What estimates are there for the amount of information a child acquires during it's acquisition of language?

Matthew: Though the probabilistic approach to parsing phrase structure looks promising, as the authors freely admit, there is considerable work left to done when moving from the simple context-free case considered by Chater & Manning to more naturalistic situations. in this regard, I would like to hear ideas & discussion on the following.

a) Are there alternative theories that are able to handle the context-free phrase structure parsing dealt with by the paper's "tree probability" model? if so, what are the strengths & weaknesses of these models relative to the tree probability model.

b) The authors note that "social & environmental context" play a role in parsing. seemingly this role is through the likelihood function. To this end a discussion of how such likelihoods may be learned & how fluid contextual change could be instantiated would be interesting.

Erin: Chater and Manning mention that there has been some work done on computationally modeling a theory of mind. Do you know of any specific work being done in this area?

II. Prior knowledge:

Erin: When Chater and Manning are discussing computational models of language learning, they say that certain aspects of language structure can be learned from the linguistic corpora by distributional methods, and they give an example of this, clustering. I'm not entirely sure how this is supposed to work, but I was wondering what sort of knowledge would this depend on? Do the learners have to know that there are noun-words, and verb-words and such in their language for this to work?

Kenny: Figure 1 raises questions about what the model literally starts out with. For example, does the model know that sentences are composed of [NP, VP] in 100% cases, giving it a foothold that this is a SVO language? Without abstract, syntactic categories, it seems like parsing might be impossible. On the other hand, it seems to me like the most interesting aspect of learning a grammar would be establishment of categories that words are binned into. I'm not even a novice on syntax theories - but do we assume that babies have categories that they are parsing speech into on the basis of experience, and that is the basis for building structure?

Pernille: I have a hard time understanding the tables in the Chater and Manning article. Can you elaborate?

2) Page 336, 2nd paragraph, line 14: I don't understand the bracketing; can you explain?

Matthew: Phrase processing is a fairly fast process, especially given the potentially large number of parses available for a given sentential structure. The tree probability model presented requires computing the all potential phrase parsings, which seems to run into problems of scalability when moving from simple examples to more naturalistic ones. From one standpoint, such scaling problems are not necessarily problematic: computing probabilities across these tree is merely the goal mental computation is trying to achieve. This standpoint seems somewhat unsatisfying. Arguably other methods of instantiating grammatical rules outside of parse trees exist, one potential candidate would be instantiating certain grammatical rules as loss functions. a discussion of such effects that various models of instantiation have on computational models of language learning & parsing, etc. would be of interest.

III. Learning

Pernille: I don't understand Chomsky's innateness argument and the poverty of the stimulus argument. Or maybe it is that I don't buy into the argument. Can you give me a better sense of Chomsky's argument?

Kenny: 1. The authors discuss using probabilistic parsing models to determine underlying what was referred to in our first class as "hypothesis space". That is, if a model is given certain information about the implicit structure of sentences a priori, then it will be able to correctly parse anything in a language it is trained upon. However, processing and parsing language in real time involves near-simultaneous and interactive representations of speech content. Is it possible to model one aspect of grammar such as syntax to determine inherent structure, while holding off on other aspects that feed into it, particularly semantic aspects?

Matthew: Please explain parse trees in greater detail.