Bayesian inference in word learning

The fundamental task of word learning is to map the phonological sound of a word to its referent or extension in the world: to realize, for instance, that the phoneme sequence /dog/ maps onto Snoopy and Fido but not the neighbor's cat. This is a difficult problem because as the philosopher W.V.O Quine (and many since) have demonstrated, the meanings of words are inherently logically under-constrained. There are an infinite number of possibilities consistent with the meaning of any word: *dog* could mean "four legged animal that barks", "undetached dog parts", "this instance of a barking animal until time *t*", "furry animals 93 million miles from the sun", and so forth. In addition, children appear to learn words gradually, and even adults do not fully agree on the exact meaning of most words. These considerations suggest that word learning is a probabilistic process. As such, one of the key theoretical frameworks in which word learning may be studied is the Bayesian paradigm.

Bayesian probability theory provides a mathematical answer to the question of how to revise beliefs in light of new data. It describes an inductively correct way of thinking which results from formalizing simple desiderata that a proper reasoning system should meet. These desiderata include propositions like: "if you see some data supporting proposition A, you should conclude that A is more plausible rather than less" and "the more you think that A is true, the less you should think it is false." Because of this formalization, Bayesian inference is optimal in the sense that a non-Bayesian reasoner attempting to predict the future will always be out-predicted by a Bayesian reasoner in the long run. Mathematically, the theory predicts which beliefs or hypotheses are best supported by some data by combining two elements: the **likelihood**, which captures the degree to which the data are predicted by the hypothesis; and the **prior**, which captures the probability of the hypotheses before having seen the data. When used in cognitive science, the prior generally captures the biases a learner brings to the situation.

The utility of Bayesian inference in understanding word learning is that it can be informative to compare the probabilistic inferences made by children and adults as they are learning words with the probabilistic inferences that Bayesian probability theory predicts a learner *should* follow. When learners behave as predicted, the theory can tell us about the reasons for their behavior; when they behave differently, the theory can be modified to explore whether the differences derive from cognitive limitations like memory or different prior biases than were originally assumed. Bayesian models has found several specific applications in word learning so far.

Inference from a few examples: the size principle. One of the central problems in word learning is understanding **fast mapping** -- the ability to acquire the meaning of a word, including identifying its full extension, given only one or a few examples. From a theoretical point of view, this is a difficult problem because we know that word learning must be probabilistic, but probabilistic inference (especially as implemented in more neurally-inspired models like connectionist networks) is often slow and gradual. Bayesian inference offers an explanation for this. Consider a scenario in which a learner has seen a single dalmatian and heard it called a "dalmatian." In this case, it is hard to know whether the word should also apply to schnauzers, the neighbor's cat, or even a bird. However, if they have been shown *three* dalmatian with that label, then most people would conclude that the word "dalmatian" does not apply to other dogs, much less cats or birds. Intuitively, this is because if it *did* apply more broadly, it would be odd indeed that one of the three examples was not something other than a dalmatian.

The mathematics of Bayesian probability theory capture this intuition formally in something called the **size principle**. This principle captures the insight that you are more likely to see a given example if it is drawn from a concept with a smaller extension. You are more likely to draw a red ball from a bag containing two balls (a red and a green) than from a bag containing five balls (a red

and four greens). In a similar way, you would be more likely to see a dalmatian offered as an example of a dalmatian than as an example of a dog, because there are many other dogs that could have been chosen in the latter case. This effect is magnified as more examples are drawn. Thus, receiving three examples of dalmatians is strong evidence that the word only refers to dalmatians, rather than schnauzers or terriers as well. Both children and adults appear to make inferences about how to generalize based on the size principle, as Bayesian theory would predict.

A key underlying assumption driving this finding is that people are sensitive to how the data are sampled from the environment -- in this case, which objects in the world are chosen as examples of a certain label. The size principle only applies if examples are sampled from the concept itself by a teacher who knows the concept. This sort of sampling, called **strong sampling**, occurs when a person specifically teaches a word (e.g., *fep*) by choosing examples of objects that are called *feps*. Another kind of possible sampling occurs when the learner -- who does not know the meaning of the word -- offers instances to the teacher to be labelled. This is called **weak sampling** and does not license the same generalizations: if I myself choose the items and they all happen to be *feps*, this does not imply that only those things I have chosen are *feps*. Indeed, both adults and children are sensitive to differences in how the data are sampled, and generalize differently in each case as Bayesian probability theory predicts.

Inference from a few examples: word learning biases. Another way in which people can generalize from just a few examples is by relying on word learning biases. For instance, consider a child who has learned that all items called by the same word tend to have the same shape -- this is called the **shape bias**. A child with the shape bias who is provided an example of a *ball* will conclude, without having to see any other balls, that the new bouncy ball (a sphere) is more likely to be a ball than the block (a cube). The Bayesian framework can explain this type of inference in two ways. Most straightforwardly, prior biases can be captured by encoding it in the prior probability. Under this approach, hypotheses about word meanings with higher prior probability would be those in which all the items called by a word share the same shape. However, this approach requires building in the bias: what about biases that might be learned, like the shape bias?

Here a kind of Bayesian model called a **Hierarchical Bayesian Model (HBM)** can provide an explanation. Normal Bayesian learning involves calculating priors and likelihoods about different hypotheses (in this case, hypotheses about the meanings of words); in HBMs learning also involves calculating priors and likelihoods about *kinds* of hypotheses. This sort of "learning to learn" may be what is going on when children acquire biases like the shape bias. The learner may be simultaneously learning about specific concepts (e.g., that balls tend to be round and bouncy, while dogs have four legs and bark) as well as forming generalizations about *kinds* of concepts (e.g., most concepts map onto things that are uniform in shape, but not texture or color). Once a bias like the shape bias is learned, fast mapping is facilitated, since the learner doesn't need multiple examples in order to know how to guess the extension of the category.

Combining multiple sources of information in word learning. Word learning does not occur in isolation. It occurs at the same time that other aspects of language, like word segmentation and grammar, are being acquired; it also occurs at the same time that other types of knowledge are being acquired, like social behavior and how objects move. In one way this joint learning problem makes the learning task harder, because there is more to acquire. But in another way it is easier, because the different areas are mutually constraining. For instance, being able to track the social and referential attention of speakers based on where they are looking may make it easier to identify the meaning of the words they are using. Bayesian models provide an excellent framework for studying this sort of learning, since multiple sources and kinds of information are all represented the same way: as probabilities about hypotheses. Moreover, Bayesian probability theory provides a

normative standard of how information from multiple sources *should* be combined in the most optimal way, and this can be compared to human performance.

One Bayesian model of learning from multiple sources of information has addressed how people might acquire a lexicon while at the same time learning about their interlocutor's referential intentions. The model makes two basic assumptions: first, that what speakers intend to say is a function of the world around them; second, that the words they utter are a function of what they intend to say and the language they are speaking. It also encodes a prior preference for a smaller lexicon: a lexicon with spurious word-object mappings that do not explain the data (which items are labeled) would not be preferred. Based on these assumptions, the model ends up preferring one-to-one lexicons, and thus developing a **mutual exclusivity bias** (a bias to think that each word has one meaning). This bias emerges because if the lexicon allows multiple words to refer to an object, then when that object occurs, the probability of any one word being said is reduced. In other words, if an object could be called a *dax,* a *plim,* or a *tref,* then the probability of hearing "dax" in the presence of the object is only 33%.

Another way that learners might combine multiple sources of information is to use information about grammar to help learn the meaning of words. A Bayesian model simultaneous learning of word order and word meaning shows that each individual learning problem is facilitated by the other, at least in a toy world. This occurs in both directions. For instance, if a learner has realized that verbs tend to come after subjects and objects, they may quickly conclude that *durk* is the action when hearing a sentence like *trif bel durk*. Conversely, knowing that *durk* refers to an action could help a learner who doesn't know the word order of the language to realize that verbs tend to come at the end.

Bayesian models of word learning have only existed since approximately 2007. As a result, there are many areas of word learning that they have not yet been applied to, and much remains to be learned about the limits of their applicability. So far Bayesian inference has been most useful as a way to understand the abstract computational task that learners are faced with when learning words -- what problems they must solve and what a solution would look like when generalizing from few examples, acquiring word-learning biases, or combining multiple sources of information. Bayesian models have also proven valuable as a normative standard against which human behavior can be compared. Ultimately, they will need to be applied to larger-scale problems and include predictions about how limitations in cognitive capabilities would change the inferences predicted. Both of these are possible within the framework; only time will tell if their promise is realized.

Amy Perfors
University of Adelaide

See also: Cross-Situational Observation in Word Learning; Induction in Language Learning; Over-Extension and Under-Extension in Word Learning; Shape Bias in Word Learning; Word Learning: Constraints

Further Readings

Frank, M.C., Goodman, N.D., and Tenenbaum, J.B. “Using speakers’ referential intentions to model early cross-situational word learning.” *Psychological Science* v.20: 579-585. (2009)

Kemp, C., Perfors, A., and Tenenbaum, J.B. “Learning overhypotheses with hierarchical Bayesian models.” *Developmental Science,* v.10: 307-321. (2007)

Maurits, L., Perfors, A., and Navarro, D.J. “Joint acquisition of word order and word reference.” *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*: 1728-1733. (2009)

Perfors, A., Tenenbaum, J.B., Griffiths, T., and Xu, F. “A tutorial introduction to Bayesian models of cognitive development.” *Cognition*, v.120: 302-321. (2011)

Quine, Willard V.O. *Word and object.* MIT Press. (1960)

Xu, F., and Tenenbaum, J.B. “Word learning as Bayesian inference.” *Psychological Review*, v.114: 245-272. (2007)