**Learning semantic structure**
NTT-MIT joint research proposal

Josh Tenenbaum
May 2002

Our project has two threads:

1) Building computational models of how people learn and structure semantic concepts, and testing those models with behavioral experiments.

2) Developing new machine learning algorithms to help computers learn and structure semantic concepts closer to the ways that people do.

Two specific questions drive this research program. First, how do people -- and how can computers -- combine unsupervised and supervised approaches to concept learning, using unsupervised learning to build better hypothesis spaces for supervised concept learning? Second, what is the large-scale structure of human semantic concepts in natural language, and how can this structure be learned – by either people or machines -- from large text or hypertext corpora? This proposal discusses each of these questions in turn. In each of these areas, there are currently ongoing efforts in both Tenenbaum's group at MIT and Ueda's group at NTT. Our overarching goal for this project is to improve each group's work by bringing it into closer contact with synergistic work in the other group, and to develop new joint approaches that combine the strengths of each group's efforts.

## I. Combining unsupervised and supervised approaches to concept learning

The specific learning task we focus on is learning the meaning of words from examples. For instance, given just a few examples of dogs all labeled with the same word "dog," the learner's goal is to infer which other objects in the world can also receive this label. This is an important task to understand from the standpoint of human cognition. It is also of long-term interest in artificial intelligence, for the purposes of increasing the fluency of human-computer interaction as well as in more constrained applications such as searching an image database to return images matching the concept in one or more examples provided by a human user.

Learning words from examples presents a more difficult inductive problem than what is currently addressed by typical computational models of supervised concept learning. Most supervised learning algorithms assume that different concepts are mutually exclusive. If each object were an instance of one and only one concept, then positive examples of one concept would necessarily be negative examples of all other concepts, and a learner who saw just a few positive examples of each concept to be learned would still have vast quantities of (primarily negative) evidence about each concept. However, natural semantic concepts are not mutually exclusive, but overlap extensively. Consider some of the overlaps with "dog": "animal", "mammal", "canine", "dalmatian", "pet", "spotted", "running", "friendly", "mean", to name just a few. To understand how people and machines can learn natural semantic concepts, we must develop new computational paradigms for learning concepts from examples.

In previous work, we have shown that human learner's generalizations of word meanings from just one or a few labeled examples can be explained in terms of Bayesian inference over a hypothesis space of possible concepts. However, we did not address the crucial question of how the learner's hypotheses might be acquired, in order to guide supervised generalization from just one or a few examples of a word.  Our present studies explore how hypotheses might be acquired for novel objects via unsupervised learning tasks from unlabeled examples.  Mixture models provide one promising framework for combining labeled and unlabeled examples -- or supervised and unsupervised learning modes -- in concept learning tasks.  Essentially, mixture models are used to discover multiple clusters in the unlabeled data, and those clusters then correspond to possible ways of generalizing a concept from a few labeled examples.

We have pursued two separate approaches in parallel at NTT and MIT.  Ueda and colleagues at NTT have adopted an approach popularized in the field of text classification (e.g., Nigam, McCallum, Thrun, and Mitchell, 2000), in which a single mixture model represents the learner's knowledge about the concepts to be learned, and labeled and unlabeled examples are processed in an integrated fashion using the EM algorithm.  They have extended the standard model in several directions, allowing for a more complex correspondence between concepts and mixture components and for more powerful learning algorithms based on variational Bayes and active learning.

Tenenbaum and colleagues at MIT have developed a new approach based on fitting multiple "one-mode" mixture models to the unlabeled data, with each model representing one potential concept to be learned from the labeled data.  Each model consists of a mixture of two components: a Gaussian distribution to capture the potential concept and a uniform noise process to account for all data points not well-described by that concept.  By fitting many instances of the one-mode mixture model to the unlabeled data, we can pick out many different clusters – or many different potential hypotheses for concepts that could be inferred from the labeled examples.  Specifically, the clusters extracted in an unsupervised fashion with the one-mode mixture models form a hypothesis space for a Bayesian (supervised) concept learning algorithm that generalizes from the few labeled examples.

Both of these approaches overcome some of the representational limitations of conventional clustering algorithms that prevent them from providing good models of the concepts picked out by words in natural language.  Conventional approaches to clustering treat clusters as either mutually exclusive or nested in a tree-like hierarchy.  In the former case, two clusters are always nonoverlapping, while in the latter case, two clusters are either completely nonoverlapping or else one is contained completely inside the other.  However, words in natural language frequently map onto partially overlapping clusters, which are neither mutually exclusive nor nested.

We have focused on two kinds of partially overlapping structures: overlapping subcategories and orthogonal classification systems.  As an example of overlapping subcategories, consider "feline" and "pet", two subcategories of "animal": a cat may be both "feline" and "pet", a dog only the latter, and a leopard is (hopefully) only the former.  As an example of an orthogonal classification system,  we might talk about the objects found in a furniture store in terms of their kind -- ``table", ``chair", ``bookcase" -- or their composition -- ``metal", ``plastic", ``wooden".  While the categories within one system (e.g., kind) are mutually exclusive, each one cross-cuts every other category in the other system (e.g., composition).

Both of our mixture-model approaches can, in different ways and to different extents, learn to represent such partially overlapping concepts from unlabeled data.  The approach of Ueda and colleagues does this by relaxing the one-to-one correspondence between concepts and mixture components that is the basis of traditional mixture models.  The approach of Tenenbaum and colleagues does this by allowing the extraction of each cluster to proceed independently of the others, so that clusters come to represent the strongest regularities in the data regardless of how they might overlap with other clusters.

From the point of view of combining unsupervised and supervised concept learning, the two approaches have complementary strengths and weaknesses. Ueda's approach is not as flexible in the kinds of cluster overlaps it can tolerate, but allows the supervised and unsupervised learning phases to be seamlessly integrated in parallel. Tenenbaum's approach allows a greater range of overlapping cluster structures, but requires that the unsupervised phase strictly precede the supervised phase, with unlabeled data influencing learning from labeled data but not vice versa. Also, Ueda's approach supports powerful variational Bayesian and active learning algorithms, while Tenenbaum's approach requires further work on search procedures for extracting all significant clusters from a data set efficiently and with minimal redundnacy. Both approaches have produced good results on several benchmark data sets, but require further testing and development on large real-world data sets.

In ongoing work, Tenenbaum, Ueda and their colleagues are discussing how to bring their proposed approaches into closer contact, through common applications or evaluations and, if possible, the development of hybrid algorithms that combine the powerful learning algorithms of Ueda's approach with the rich representational possibilities of Tenenbaum's approach. We are focusing on tasks that people perform relatively effortlessly and successfully, far beyond the capabilities of existing machine learning algorithms . In particular, we are exploring the extent to which our models can explain the behavior of human learners in clustering, sorting, and word learning tasks, and we are working on a tool for automated conceptual image-database search that attempts to use the huge volumes of unlabeled data available in many databases to infer more accurately what kinds of results a human user desires from a query that consists of just one or a few labeled examples.

## II. Learning semantic structure from text and hypertext

The work described above focuses on learning single concepts, one at a time, primarily from perceptual input. However, many concepts – particularly more abstract concepts – are probably learned in a very different way: by encountering them in one or more linguistic contexts (documents) and integrating them into a large network of previously learned concepts. In order to understand more fully these aspects of human concept learning, we have recently turned our attention to studying the large-scale structure of semantic networks in natural language. We have focused on trying to describe this structure statistically, and also on developing learning algorithms that can extract this structure from large text or hypertext corpora and put it to use in information retrieval tasks.

### A. Statistical analyses of network structure

Our statistical studies of the large-scale structure of semantic networks have focused on networks of word associations obtained from three different sources: human subjects in free association experiments, Roget’s thesaurus, and WordNet. We have found several statistical properties in common across these networks. First, they all have a small-world structure, characterized by sparse connectivity, short average path-lengths between words, and strong local clustering. In addition, the distributions of the number of connections follow power laws that indicate a scale-free pattern of connectivity, with most nodes having relatively few connections joined together through a small number of hubs with many connections. These regularities have also been found in certain other complex natural networks, such as the world wide web, but they are not consistent with many conventional models of semantic organization, based on inheritance hierarchies, arbitrarily structured networks, or high-dimensional vector spaces. We have proposed that these structures reflect the mechanisms by which semantic networks grow. We have also developed a simple model for semantic growth, in which each new word or concept is connected to an existing network by differentiating the connectivity pattern of an existing node. This model generates appropriate small-world statistics and power-law connectivity distributions, and also suggests one possible mechanistic basis for the effects of learning history variables (age-of-acquisition, usage frequency) on behavioral performance in semantic processing tasks.

In future work, we would like to better understand the implications of these statistical properties of natural semantic networks, and the growth processes giving rise to them, for processes of information retrieval and search. Algorithms for efficient search on small-world and scale-free networks have been developed, which when coupled with our semantic network representations, could lead to better models of human semantic memory retrieval as well as more efficient computational models for constructing coherent semantic interpretations of sentences.
Also, we would like to draw connections between the model of network growth developed in Tenenbaum's group and a new model recently proposed by Saito, Ueda, and colleagues at NTT. The NTT model was proposed to describe the web, with an explicit focus on community structure that gives rise to strong clustering, and an elegant probabilistic formulation in terms of latent variables. We would like to explore the community structure of semantic networks, with the expectation that communities will be found corresponding to coherent semantic domains or topics. We would also like to explore whether the latent-variable formulation of the NTT model can be profitably combined with the growth mechanisms of our previously developed model for semantic growth, to produce more realistic models of natural-language semantic networks. Finally, we would like to investigate the possibility of creating joint models of semantic networks and web structures, which would mutually constrain each other: knowledge of web community structure would lead to more accurate semantic networks, and knowledge of semantic structure would lead to more accurate models of web communities.

B. Algorithms for learning semantic networks

The statistical patterns described above place strong constraints on models that attempt to learn semantic concepts from large text corpora. We have shown that popular methods for information retrieval such as latent semantic analysis (LSA) are not capable of producing these patterns. This has motivated us to develop new approaches for learning semantic structure from text.

One approach is based on Latent Dirichlet Allocation, recently introduced by Blei, Ng, and Jordan (2001). We have developed a Markov Chain Monte Carlo approach to fitting this model, and shown that it is capable of learning meaningful semantic representations from raw text. We have also shown that it produces representations consistent with the statistics of word associations in natural language, as described above. That is, the distribution of word senses follows a power law, and the network's connectivity behaves as a small world. This model is the first that we know of that can learn, from raw text, meaningful semantic representations with large-scale statistical properties that are qualitatively similar to natural language semantic networks.

Recently, Ueda and Saito at NTT have developed a new approach to learning multiple classifications in the context of classifying text documents by topic. Their PMM algorithm is designed to learn document classifications which are not necessarily mutually exclusive – a critical goal if these classes are to correspond to natural language concepts, as explained above. They have shown that this algorithm outperforms conventional approaches based on binary classification, when applied to real data from the WWW. The approach of Ueda and Saito is closely related to both the one-mode mixture model and the latent dirichlet model that have been explored in Tenenbaum's group. A high priority for future work is to pursue the connections between these models and the PMM approach developed at NTT. We hope to produce a joint paper soon applying these methods to two related problems in learning from text: classifying documents into overlapping concepts (topics), and classifying words into overlapping concepts (senses). Over the longer term, we also hope to bring these approaches to bear on the other main theme of this project: combining supervised and unsupervised concept learning. Because these approaches are formulated as probabilistic mixture models, they provide a natural framework for unsupervised learning of hypothesis spaces that could be useful in subsequent supervised concept learning. Given their success on learning from raw text, it would be most interesting to see if they can also be adapted to learn from a combination of text and perceptual input, and thus provide a more general framework for understanding the different aspects of human concept learning.

## III. Approximate budget

1 PhD student RA: $50,000 (total cost)
1 Technical Assistant 50% time: $30,000 (based on full time 1-year salary $28K)
1 month summer salary for Tenenbaum: $15,000 (based on 2001-2002, 9 month salary of $65K)
Computer Equipment (2 workstations): $12,000
Travel and miscellaneous: $3,000