Theory and Application of Concept Learning
MIT2002-01

Progress Report: July 1, 2002—December 31, 2002

Joshua Tenenbaum

Project Overview

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. 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?

Progress Through December 2002

We have made progress on several computational models of human semantic knowledge and learning. These models draw on – and also have the potential to advance – the state of the art in machine learning and statistics. They also have the potential for significant applications in information retrieval.

One model addresses the problem of learning new concepts from just a few examples, in a domain where the examples have some meaningful, previously learned semantic structure. Our current domain focus is biology. The instances are kinds of animals, e.g. “gorillas”, “horses”, or “squirrels”, and the concepts are biological predicates, e.g., “can get the disease HAV” or “has nicotinic acid in the blood”. Example tasks include *specific* queries – e.g., given that horses and squirrels can both get the disease HAV, how likely is it that gorillas can get HAV? – as well as *general* queries – given the same examples, how likely is it that all mammals can get HAV? Tasks such as these are essential for common-sense reasoning. We have developed a Bayesian model that assumes concepts can be represented as disjunctive combinations of one or more clusters from a tree-structured taxonomy, with a prior that penalizes more complex hypotheses (consisting of more disjuncts). This model accounts for a number of qualitative phenomena of human inference in this domain, and also fits human judgments quantitatively to a high degree of accuracy (with only one free parameter controlling the magnitude of the complexity penalty). The tree is built by hierarchical clustering on the pairwise similarities between instances, but could also constructed by clustering on high-dimensional vectors describing the unlabeled data, or be hand-designed based on domain knowledge. This model thus offers a principled general-purpose approach to learning concepts from a combination of very few labeled examples and prior domain knowledge (specified either explicitly in the form of a tree-structured taxonomy or implicitly in the similarity metric or distribution of unlabeled data). It is described in Sanjana and Tenenbaum (2003), one of four finalists for the Best Student Paper prize at NIPS 15 in Dec. 2002.

Another model addresses the problems of learning semantic structure from large text corpora and relating that structure to the human sense of semantic association between words. This model is based on the Latent Dirichlet Allocation (LDA) model of Blei, Ng, and Jordan (2002), which has been shown to learn a basis of semantically meaningful topics for representing and retrieving text documents, and which is related to the Parametric Mixture

Model (PMM) developed by Ueda-san's group at NTT CS Labs. Our model has two modifications over the basic LDA model. First, we embed the LDA model as one state of a hidden markov model (HMM), in which other states capture the statistics of transitions between words that reflect general syntactic regularities, independent of topic-specific semantic structure. Second, we train the model using a simple Markov chain Monte Carlo (MCMC) procedure. The results are (1) a separation of primarily semantic and primarily syntactic words (or word senses) into the topics of the LDA model and the alternative states of the HMM, respectively; and (2) purer semantic topics, without the need for an arbitrarily chosen "stop list". We believe this method may be the first to discover the syntax-semantics factorization in a purely bottom-up fashion, assuming only that there are distinct components of the representation specialized for topic-specific statistics independent of word order and word-order statistics independent of topic. We have also shown that this model's estimates of conditional probability $p(w_i|w_j)$ – the probability that word *i* will occur in a document given that word *j* occurs in the document – provide a much better model of human word association data than alternative models based on Latent Semantic Analysis (LSA). In particular, word choices in free association, as well as power laws in the distribution of the number of associates per word, are better captured by our model. These results are described in Griffiths and Steyvers (2002) and Griffiths and Steyvers (2003).

Research Plan for the Next Six Months

In the next six months, we plan to pursue a range of potential applications for these two models. We have discussed all of these plans with Ueda-san's group at NTT CS Labs, and we are hoping to collaborate on several of them. For the Bayesian concept learning model, we intend to explore content recommendation in domains with naturally tree-structured taxonomies. This includes large-scale catalogs, e.g. Amazon.com, as well as the web itself, as indexed by the open directory project. A sample application might be to take the links on a person's homepage as a random sample of all the pages relevant to them, and then to rank all other sites in the open directory in terms of their estimated probability of being relevant to the person. Applying our model to a large data set will require developing approximations to the full Bayesian sum over all logically possible hypotheses. We hope to exploit the tree-structured constraints on the prior to make these approximations efficient and to bound their accuracy theoretically.

For the LDA-HMM model, we intend to explore applications to information retrieval and text classification with very small text samples, e.g. individual sentences rather than full documents. We expect that the factorization of syntactic and semantic structures provided by our model will offer greatest benefits in this setting. Pursuing this application will require developing efficient approximate methods for inference in the LDA-HMM model. Currently, the loopy structure of this model makes standard belief propagation impractical. In the near term we will be investigating variational methods (in collaboration with David Blei) and expectation-propagation methods (in collaboration with John Lafferty).

Finally, we are pursuing two new directions (again, in collaboration with Ueda-san's group at NTT CS Labs). First, we are developing new models that extend the approaches described above to the problem of learning the perceptual grounding of words, based on a large database of images of objects plus associated words. This problem domain is challenging because different kinds of words refer to different kinds of image structure, e.g. shape, color, texture, which do not simply correspond to a single taxonomic hierarchy. We expect that it will provide a productive domain for developing more powerful approaches to learning concepts from few labeled examples plus prior domain knowledge. Second, we are developing novel approaches for embedding network structures, such as networks of semantic associations, in Euclidean space. These techniques are suitable for revealing the intrinsically meaningful structures in these networks, even if their metric structure (defined in terms of geodesic paths between nodes) is highly non-Euclidean.

References

D. M. Blei, A. Y. Ng, and M. I. Jordan (2002). Latent Dirichlet Allocation. *Journal of Machine Learning Research* , Vol. 3, pp. 993-1022, 2003.

T. Griffiths and M. Steyvers (2002). A probabilistic approach to semantic representation. In *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*. Erlbaum.

T. Griffiths and M. Steyvers (2003). Prediction and semantic association. In *Advances in Neural Information Processing Systems 15*. MIT Press.

N. Sanjana and J. B. Tenenbaum (2003). Bayesian models of inductive generalization. In *Advances in Neural Information Processing Systems 15*. MIT Press.