Artificial General Intelligence through Large-Scale, Multimodal Bayesian Learning

Brian MILCH [a,1],

[a] *CSAIL, Massachusetts Institute of Technology, USA*

Abstract. An artificial system that achieves human-level performance on open-domain tasks must have a huge amount of knowledge about the world. We argue that the most feasible way to construct such a system is to let it learn from the large collections of text, images, and video that are available online. More specifically, the system should use a Bayesian probability model to construct hypotheses about both specific objects and events, and general patterns that explain the observed data.

Keywords. probabilistic model, architecture, knowledge acquisition

Introduction

A long-standing goal of artificial intelligence is to build a single system that can answer questions as diverse as, “How can I get from Boston to New Haven without a car?”, “How many Nobel Prizes have been won by people from developing countries?”, and “In this scene showing people on a city street, about how cold is it?”. Answering such questions requires broad knowledge, on topics ranging from public transit to geography to weather-appropriate clothing. These questions also require deep reasoning, not just scanning for keywords or looking at simple features in an image.

So far, we do not have AI systems whose knowledge is both broad and deep enough to answer this range of questions. The most prominent efforts to acquire such knowledge are Cyc [1] and Open Mind [2], both of which have significant limitations. The knowledge they collect is primarily deterministic: it does not include quantitative measures of how often things tend to occur. Furthermore, adding new knowledge requires effort by humans, which limits the breadth of questions that can be answered.

Meanwhile, other branches of AI have focused on reasoning with probabilistic models that explicitly quantify uncertainty [3], and on learning probabilistic models automatically from data [4]. This probabilistic approach to AI has been successful in narrow domains, ranging from gene expression analysis [5] to terrain modeling for autonomous driving [6]. It has also seen domain-independent applications, such as sentence parsing [7] and object recognition [8], but these applications have been relatively shallow: they have not captured enough semantics to answer questions of the kind we posed above.

[1] Corresponding Author: Brian Milch, MIT CSAIL, 32 Vassar St. Room 32-G480, Cambridge, MA 02139, USA; E-mail: milch@csail.mit.edu.

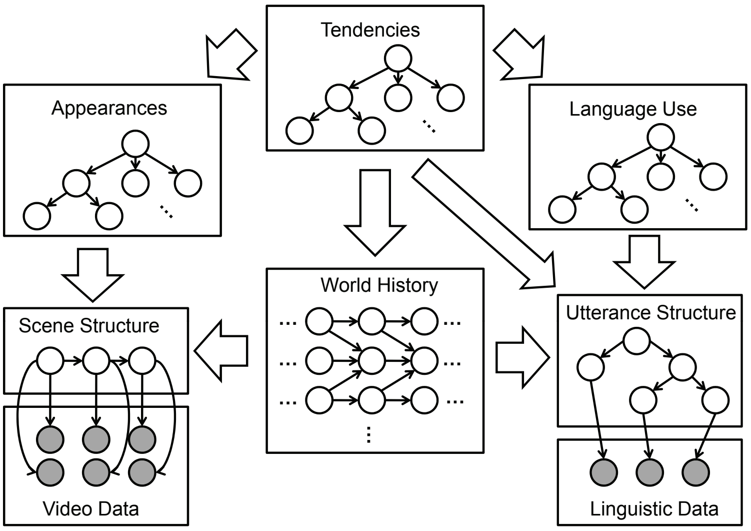

Figure 1. Sketch of a probabilistic graphical model for AGI. Nodes represent random variables and arrows represent probabilistic influences.

Thus, for the past two decades there has been an unfortunate divide between the probabilistic AI community and those actively working on deep, open-domain AI — what Goertzel has called *artificial general intelligence* or *AGI* [9]. There are some signs that this gap is narrowing: the Novamente Cognition Engine [9] is an AGI system that does probabilistic reasoning, and the probabilistic AI community is showing renewed interest in large-scale knowledge acquisition [10,11].

The purpose of this paper is to bridge this gap more completely. We propose a Bayesian AGI system that takes a large chunk of the web as input, including both hypertext and video data, and uses a probabilistic model to make inferences about what the world must be like. We also comment on integrating AGI into the ecosystem of probabilistic AI research.

1. A Bayesian Architecture for AGI

Figure 1 shows a schematic view of a probabilistic model suitable for an AGI system. The shaded nodes represent variables that the system can observe: pixel values in videos, and words in linguistic data. The unshaded nodes represent unobserved variables about which the system can form hypotheses. The variables in the top box represent general tendencies in the world, such as the way public transit tends to work, and the tendency of people to wear wool hats and scarves only in cold weather. The “world history” box contains variables representing specific facts about the world: for instance, the current public transit schedules between New Haven and Boston, and the history of Nobel prizes. The large arrow from the “tendencies” box to “world history” indicates that general tendencies influence the concrete facts. There are also arrows within the world history box, representing causal links between events and their successsors.

In the top left and top right boxes, theare are variables for how things tend to appear, and how people tend to use language. Finally, there are variables representing the 3-d structures that underlie particular scenes and the syntactic structures that underlie particular utterances. These variables depend on the facts in the world: a scene or utterance typically provides a window on some piece of the world history. The additional arrow from tendencies to utterance structure allows utterances to be directly about tendencies: for instance, someone could say, “People wear scarves when it’s cold”. We discuss parts of this model in more detail in Sections 3 and 4.

Our system’s probability model can be seen as a joint distribution over all thse variables. It can also be understood as a hierarchy of models: a hypothesis about the variables in the top three boxes defines the structure and parameters of a probability model for world histories, scenes, and utterances. In using a top-level probability model to define a distribution over sub-models, we are taking a Bayesian approach [12].

Bayesian modeling is attractive because it makes an automatic trade-off between the complexity of a hypothesized sub-model and how well it fits the data. Complex hypotheses may fit the data better, but they typically have lower prior probabilities than simple ones. Furthermore, hypotheses involving many free parameters are penalized by a Bayesian Ockham’s razor effect [13]: it is unlikely that all those parameters would happen to have values that match the data. Although similar trade-offs arise from Kolmogorov complexity [14] and the minimum description length principle [15], the Bayesian approach lets us express prior knowledge about the structure of the world, and also maintain uncertainty about sub-models rather than choosing a single best hypothesis. There are also close connections with Bayesian approaches in statistics [12], epistemology [16] and cognitive science [17].

2. Training Data

There are several ways in which one could hope to train a Bayesian AGI system. One is to train it solely on text, a huge amount of which is available online. However, this would leave the system unable to answer queries about images or video scenes. Furthermore, there is no evidence that it is possible to build a deep model of the world from lingusitic input alone: human children all have access to visual, or at least tactile, input.

Alternatively, we might train our AGI system by hooking it up to one or more mobile robots. This would have several advantages: the system could learn from continuous streams of experience with multiple sensors, and it could actively move around and manipulate its environment to reduce its uncertainty. On the other hand, a robot’s experience would be fairly narrow, limited to the environments where its owners let it move around.

Fortunately, given the explosive growth of online video, we no longer have to choose between a large, diverse data set and a multimodal one. Videos available online include many everday objects and activities (sometimes in the background), often correlated with speech. Meanwhile, the text on the web covers many kinds of objects and phenomena that are not readily visible in videos. Thus, we propose to train our AGI system on a web data set that includes hypertext, images, and video.

In contrast to a human child or a robot, this system still will not have a chance to actively explore its world. But acquiring multimodal data from the web promises to be easier and less expensive than acquiring it with a fleet of mobile robots — and even human children learn a lot from non-interactive media such as books, television, and movies.

3. Built-In Model Components

Although Figure 1 provides a rough sketch, the question of how to define the probability model for a Bayesian AGI system is still largely open. Some authors have argued that we should be able to use generic models that encode little or no prior knowledge about the physical world [18]. But finding good hypotheses to explain the world is a daunting search problem; to the extent that we can help our system along, it seems that we should do so. After all, it is unlikely that human children start from scratch: we seem to have an evolved infrastructure for intepreting visual input, understanding language, and interacting with other people.

Modeling the physical world is one area where we should be able to provide our AGI system with a good running start. The computer graphics and animation communities have done a great deal of work on modeling the world with enough fidelity to create compelling visual environments. Our appearance sub-models should be able to use standard mesh data structures for representing the 3-d shapes of objects, and also borrow skeleton-based techniques for modeling articulated objects. The system will still need to infer the meshes and skeletons of objects (and classes of objects) from data, but at least it will not need to invent its own approach to representing shape and movement.

It also makes sense to build in components for dealing with speech, hypertext, and language in general. The language-use sub-model should contain built-in representations for words and parse trees. As with the appearance model, all the specific content — the grammars of particular languages, and the words and phrases people tend to use for particular meanings — should be inferred by the system. But there is no reason for the system to re-invent the basic ideas of computational linguistics.

Another area that calls out for built-in infrastructure is reasoning about the beliefs and desires of other agents. This ability is crucial both for understanding human behavior in the training data, and for giving helpful answers to users' questions. The system must be able to put itself in another agent's shoes, using its own inference and planning capabilities to reason about what another agent probably believes and what it might do to achieve its goals. Thus the system's probability model must have some characteristics of a probabilistic epistemic logic theory [19] or a multi-agent influence diagram [20]. It is particularly hard to imagine how a Bayesian AGI system might invent this intentional view of other agents if it were not built in.

4. Learned Model Components

Of course, even with built-in capacities for physical, linguistic, and multi-agent reasoning, our system will have a lot to learn. It will lack such basic notions as eating, smiling, driving, owning objects, having a disease, and so on. Thus, the overall probability model must admit indefinitely many non-built-in concepts.

Fortunately, there has recently been a surge in research on probability models with an unbounded number of latent concepts, using a non-parametric Bayesian construct called the Dirichlet process (DP) mixture model [21,22,23]. A DP mixture can be thought of as a graphical model where the nodes are indexed by an infinite collection of objects, and these latent objects are "recruited" as necessary to explain the data. The Bayesian Ockham's razor effect [13] prevents the model from hypothesizing a separate latent object for each data point. DP mixtures can serve as building blocks in quite complicated models, including those that model the relations among objects as well as their properties [24].

In addition to hypothesizing new concepts, our system must learn how they relate to one another. Often, these dependencies are probabilistic: for example, a person's clothing choices depend probabilistically on the weather, and a whether a person knows French depends probabilistically on where that person grew up. There has been considerable work on learning the dependency structure of probabilistic graphical models [25]. However, just learning dependencies among individual variables — Laura's clothing choices depend on the weather in New York, Mary's choices depend on the weather in London, etc. — is not sufficient. Instead, our system must learn dependency models that apply to whole classes of objects, specifying how probabilistic dependencies arise from underlying relationships (such as the relationship between a person and the city where she lives). Learning such dependency models is the domain of a young field called statistical relational learning [26]. Considerable progress has been made on learning certain classes of relational dependencies [27,28], although learning models in highly expressive languages such as Bayesian logic [29] remains an open problem.

5. Algorithms

Because we are taking a Bayesian approach, our proposed AGI system does make a strict separation between reasoning and learning: it learns by probabilistic inference. But of course, probabilistic inference can be very computationally expensive (a naive approach to answering a query would involve summing or integrating over all the unobserved, non-query variables). We have no silver bullet for this problem, but there are indications that performing inference on an AGI model is not as ridiculous as it may sound.

The first thing to note is that we will be satisfied with approximate inference methods, even if they do not provide bounds on the quality of the approximation — we will be able to judge if we are doing well enough based on how well the system answers queries. The major approaches to approximate inference today are Markov chain Monte Carlo (MCMC) [30], mean-field-based variational methods [31,32], and belief propagation methods [33,34]. These methods can take hours to run on sets of just a few hundred documents or images, so one might object that using them on a large chunk of the web is infeasible. But given a hypothesis about the world history and the appearance and language-use sub-models, the individual videos and documents in our training set are conditionally independent of each other. Thus, the work of parsing these data items can be parallelized across a large set of machines.

Perhaps the greatest challenge for inference in an AGI model is preventing the system from considering too many variables when evaluating a hypothesis or answering a query. For instance, consider trying to judge the ambient temperature in a street scene showing just one man, who is wearing a winter coat but no hat or scarf. In principle, a multitude of variables are relevant to judging the temperature: Is this person just stepping outside briefly, or is it warm enough that he can take a long walk in these clothes? What is his reason for being outside? Is he a local, or is he visiting from someplace warmer or colder? These variables are all within the purview of an AGI system — and indeed, with different data or a different query, it might be important to reason about them. But in this case, reasoning about them is bound to be unproductive: the system will not be able to conclude anything strong enough about these variables to influence the query result. Thus, the inference algorithm must be able to deal with *partial* hypotheses, and sum out the remaining random variables without reasoning about them explicitly. Existing algorithms can exploit situations where certain variables are strictly irrelevant given the hypothesized values of other variables [35]. There are also algorithms that exploit interchangeability among objects to sum out large sets of variables in a single step [36,37]. But dealing with queries like the one above will require novel methods for exploiting approximate irrelevance and approximate symmetry among variables.

Another relevant line of research is the integration of probabilistic inference with logical reasoning. Many dependencies are nearly deterministic, and our system should be able to exploit near-determinism to speed up reasoning (by contrast, standard approximate inference algorithms tend to get stuck in local optima when determinism is present). One approach is to reduce probabilistic inference to weighted model-counting in propositional logic, and then exploit fast, deterministic SAT algorithms [38]. Another approach combines MCMC with SAT algorithms based on stochastic local search [39]. But for a full-blown AGI system, new inference algorithms are sure to be necessary. One positive side-effect of working with a theoretically simple Bayesian model is that the resulting inference problems can serve as fodder for the approximate inference community. It should even be possible to extract sub-networks from the AGI model to use as highly relevant benchmark problems.

6. Measures of Progress

In order for Bayesian AGI to be viable as a research program, there must be some ways of evaluating its progress — short of seeing it pass a Turing test. One advantage of training on open-domain web data, as opposed to in an artificial or restricted environment, is that the system should quickly gain the ability to at least *attempt* open-domain tasks of practical interest. Several such tasks are used for evaluation in the computer vision and natural language fields, including the CalTech-256 object recognition data set [8], the MUC-6 noun phrase coreference data set[2], the TREC question answering evaluation [40], and the PASCAL textual entailment challenge [41]. It is unlikely that the AGI system will immediately beat state-of-the-art but shallower approaches to these tasks. However, we should at least be able to measure our system's performance and see it improving year by year.

Another way in which a Bayesian AGI system could succeed early on is by becoming a resource for shallower but more finely tuned systems. For instance, some current natural language systems use WordNet [42] or Cyc [1] to define features for machine learning algorithms, or to implement one part of a multi-stage heuristic system. If our AGI system came to play a similar role and improved the performance of the systems that used it, we would be providing a useful service. Eventually, of course, we would hope for the AGI system to surpass shallower systems that used it as a resource.

[2] http://cs.nyu.edu/cs/faculty/grishman/muc6.html

7. Conclusion

We have argued that Bayesian learning from large quantities of multimodal web data is the most feasible path toward artificial general intelligence. This position involves several strong claims: that an AGI system must learn most of its knowledge but should exploit built-in subsystems for visual, linguistic, and multi-agent reasoning; that learning from text alone is insufficient but text plus video is enough; and that Bayesianism provides an adequate theoretical foundation for AGI. These claims will be put to the test in a Bayesian AGI project.

This paper is by no means a complete design for an AGI system. The process of designing the probabilistic model, and algorithms to do inference on it, promises to involve much trial and error over a decade or two. But there seems to be a good chance of success. Furthermore, the existence of such a project (or more than one!) would be a boon to the probabilistic AI community, serving as a source of motivating problems and a testbed for new techniques. A flourishing Bayesian AGI project would bridge the disturbing gap between the successes of probabilistic AI and the goal of understanding deep, general intelligence.

References

[1] D. B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[2] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu. Open Mind Common Sense: Knowledge acquisition from the general public. In *Proc. 1st Int'l Conf. on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*, 2002.

[3] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, revised edition, 1988.

[4] M. I. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1998.

[5] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller. Rich probabilistic models for gene expression. *Bioinformatics*, 17(Suppl 1):S243–252, 2003.

[6] S. Thrun, M. Montemerlo, and A. Aron. Probabilistic terrain analysis for high-speed desert driving. In *Proc. Robotics Science and Systems Conf.*, 2006.

[7] M. Collins. Three generative, lexicalised models for statistical parsing. In *Proc. 35th Annual Meeting of the Assoc. for Computational Linguistics*, 1997.

[8] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[9] B. Goertzel. Patterns, hypergraphs and embodied general intelligence. In *IEEE World Congress on Computational Intelligence: Panel Discussion on "A Roadmap to Human-Level Intelligence"*, 2006.

[10] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the Web. In *Proc. 20th Int'l Conf. on Artificial Intelligence*, 2007.

[11] W. Pentney, M. Philipose, J. Bilmes, and H. Kautz. Learning large-scale common sense models of everyday life. In *Proc. 22nd AAAI Conf. on Artificial Intelligence*, 2007.

[12] C. P. Robert. *The Bayesian Choice*. Springer, New York, 2nd edition, 2001.

[13] E. T. Jaynes. Inference, method and decision: Towards a Bayesian philosophy of science (book review). *J. Amer. Stat. Assoc.*, 74(367):740–741, 1979.

[14] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 2nd edition, 1997.

[15] J. Rissanen. *Information and Complexity in Statistical Modeling*. Springer, 2007.

[16] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*. Open Court, La Salle, IL, 1989.

[17] A. Gopnik and J. B. Tenenbaum. Bayesian networks, Bayesian learning, and cognitive development. *Developmental Science*, 10(3):281–287, 2007.

[18] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, 2005.

[19] R. Fagin and J. Y. Halpern. Reasoning about knowledge and probability. *J. Assoc. for Computing Machinery*, 41(2):340–367, 1994.

[20] D. Koller and B. Milch. Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior*, 45(1):181–221, 2003.

[21] T. S. Ferguson. Bayesian density estimation by mixtures of normal distributions. In M. H. Rizvi, J. S. Rustagi, and D. Siegmund, editors, *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*, pages 287–302. Academic Press, New York, 1983.

[22] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *J. Computational and Graphical Statistics*, 9:249–265, 2000.

[23] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 101:1566–1581, 2006.

[24] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proc. 21st AAAI National Conf. on Artificial Intelligence*, 2006.

[25] D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[26] L. Getoor and B. Taskar, editors. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[27] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *Proc. 16th International Joint Conference on Artificial Intelligence*, pages 1300–1307, 1999.

[28] S. Kok and P. Domingos. Learning the structure of Markov logic networks. In *Proc. 22nd International Conf. on Machine Learning*, pages 441–448, 2005.

[29] B. Milch, B. Marthi, S. Russell, D. Sontag, D. L. Ong, and A. Kolobov. BLOG: Probabilistic models with unknown objects. In *Proc. 19th International Joint Conference on Artificial Intelligence*, pages 1352–1359, 2005.

[30] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, 1996.

[31] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[32] E. P. Xing, M. I. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Proc. 19th Conf. on Uncertainty in Artificial Intelligence*, pages 583–591, 2003.

[33] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proc. 15th Conf. on Uncertainty in Artificial Intelligence*, pages 467–475, 1999.

[34] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems 13*, pages 689–695. MIT Press, Cambridge, MA, 2001.

[35] B. Milch and S. Russell. General-purpose MCMC inference over relational structures. In *Proc. 22nd Conf. on Uncertainty in Artificial Intelligence*, pages 349–358, 2006.

[36] D. Poole. First-order probabilistic inference. In *Proc. 18th International Joint Conference on Artificial Intelligence*, pages 985–991, 2003.

[37] R. de Salvo Braz, E. Amir, and D. Roth. Lifted first-order probabilistic inference. In *Proc. 19th International Joint Conference on Artificial Intelligence*, pages 1319–1325, 2005.

[38] T. Sang, P. Beame, and H. Kautz. Performing Bayesian inference by weighted model counting. In *Proc. 20th AAAI National Conf. on Artificial Intelligence*, pages 475–482, 2005.

[39] H. Poon and P. Domingos. Sound and efficient inference with probabilistic and deterministic dependencies. In *Proc. 21st AAAI National Conf. on Artificial Intelligence*, pages 458–463, 2006.

[40] E. M. Voorhees. Overview of TREC 2004. In *The Thirteenth Text Retrieval Conference*. NIST, 2004.

[41] I. Dagan, O. Glickman, and B. Magnini. The PASCAL recognising textual entailment challenge. In *Proc. PASCAL Challenges Workshop on Recognising Textual Entailment*, 2005.

[42] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.