Foundations of Software Technology and Theoretical Computer Science (Bangalore) 2008.
Editors: R. Hariharan, M. Mukund, V. Vinay; pp -

Knowledge Infusion: In Pursuit of Robustness in Artificial Intelligence*

Leslie G. Valiant
Harvard University
valiant@seas.harvard.edu

ABSTRACT. Endowing computers with the ability to apply commonsense knowledge with human-level performance is a primary challenge for computer science, comparable in importance to past great challenges in other fields of science such as the sequencing of the human genome. The right approach to this problem is still under debate. Here we shall discuss and attempt to justify one approach, that of *knowledge infusion*. This approach is based on the view that the fundamental objective that needs to be achieved is *robustness* in the following sense: a framework is needed in which a computer system can represent pieces of knowledge about the world, each piece having some uncertainty, and the interactions among the pieces having even more uncertainty, such that the system can nevertheless reason from these pieces so that the uncertainties in its conclusions are at least controlled. In knowledge infusion rules are learned from the world in a principled way so that subsequent reasoning using these rules will also be principled, and subject only to errors that can be bounded in terms of the inverse of the effort invested in the learning process.

1 Introduction

One of the most important challenges for computer science is that of understanding how systems that acquire and manipulate commonsense knowledge can be created. By commonsense knowledge we mean knowledge of the kind that humans can successfully manipulate but for which no systematic theory is known. For example, conducting appropriate everyday conversations among humans requires such commonsense knowledge, while the prediction of the trajectory of a projectile can be accomplished using the systematic theory offered by physics.

We argue that to face this challenge one first needs a framework in which inductive learning and logical reasoning can be both expressed and their different natures reconciled. The learning provides the necessary robustness to the uncertainties of the world. It enables a system to go to the world for as much data as needed to resolve uncertainties. The reasoning is needed to provide a principled basis for manipulating and reaching conclusions from the uncertain knowledge that has been learned. The process by which we can infuse a system with commonsense knowledge, in a form suitable for such reasoning, we call *knowledge infusion* or *KI* [13, 15].

Robust logic [14] is a concrete proposal for realizing KI. It offers a formalism for learning rules that are suitable for later chaining together for the purpose of reasoning. In this system both learning and reasoning can be performed in polynomial time, and, further, the reasoning has certain soundness and completeness properties, and the errors in learning

*This work was supported in part by grant NSF-CCF-04-27129.

and reasoning can be upper bounded in terms of the inverse of a polynomial function of the effort expended in the learning.

For brevity we shall refer to a system that can successfully reason with commonsense knowledge as an *intelligent system*. Recent headlines in the New York Times - with one word omitted in each case - included "Can Weeds Help Solve the ..... Crisis?" and "Oil Hits New High as Dow Flirts With ..... Territory." It would be reasonable to expect intelligent systems to be able to make reasonable guesses of the missing words. To achieve such capabilities there is a need both for access to commonsense knowledge, as well as for an ability to apply such knowledge to situations not previously experienced.

We suggest that for such a word completion, or any other, task to be a valid test for intelligent systems, it will need to have two properties in common with the Turing Test [11]. First, there should be no *a priori* restrictions, to any limited subdomain or microworld, on the domain of knowledge treated. Second, there needs to be some numerical evaluation of performance relative to some baseline. We regard these two properties as the most fundamental prerequisites for tests of progress in this area.

Recently we have reported on the results of experiments that test whether KI is effective for such an unrestricted word completion task [8]. These experiments, performed on a data set of a half a million natural language sentences, showed that this task of predicting a deleted word from a sentence could be performed to a higher accuracy by this method than by a baseline learning method that did not use reasoning. In this experiment the learned rules contained commonsense knowledge about the world, while the baseline method could be regarded as a more syntactic learning method, in the sense of n-gram methods in natural language processing but using more powerful Winnow based learning methods as developed by Roth and his coworkers [4]. The experiments highlight the technical challenges of learning from noisy data reliably enough that the learned rules could be chained together fruitfully. In particular there is a need for algorithms that have good run times and good generalization properties, and for methods of chaining rules that preserve the generalization guarantees.

Technical descriptions of the approach described can be found in references [8, 14, 15] and we shall not detail any of that here. In this note we shall attempt to summarize informally the general justification of our approach in comparison with some alternatives. Since the effort needed to endow computer systems with usable commonsense knowledge can be expected to be very considerable, it seems worthwhile to invest effort into evaluating carefully the various available approaches.

2 Achieving Robustness

As soon as the feasibility of large scale computations became evident in the middle of the twentieth century an immediate concern was whether the execution of millions of instructions, each one highly accurate in itself, would inevitably lead to errors accumulating and giving totally incorrect final answers. For processing commonsense knowledge this robustness problem would appear to be an especially important concern, since significant uncertainties may appear here even in individual steps. This paper is predicated on the proposition that any theory of commonsense reasoning that fails to include robustness in its subject

matter will also fail as a basis for intelligent systems as these scale up.

That *some* theoretical basis is required for intelligent systems to be successfully realized is widely acknowledged. The system is expected to make determinations for circumstances that may not be foreseen by the designer, and these determinations will need therefore to be derived using some principled basis. In this pursuit the most widely advocated theories have been the equivalents of the predicate calculus [6], on the one hand, and Bayesian reasoning [10], on the other, or some combination of these. These theories do have much to offer, being mathematically consistent theories that attempt to address directly the representation of knowledge. Their main drawback from our perspective is that they do not address directly the issue of robustness. Indeed, as a broad generalization, it has proved in practice that systems based on these theories are brittle, in the sense that, as the knowledge bases in such systems grow, the predictions made by them degrade significantly. This phenomenon is not difficult to explain. These theories guarantee accuracy of predictions only if the model created in terms of them is consistent and accurate. Such guarantees of accuracy and consistency are not available in areas, such as commonsense knowledge, which we define here to be just those for which no exact axiomatization is known.

We regard the Bayesian framework as an elaboration of the classical logical one. It is appropriate in cases where the knowledge being axiomatized contains probabilistic processes, and there is some hope of an axiomatization. Since it is just an elaboration of logic, and in that sense at least as difficult to apply to model complex knowledge, we do not regard it as helpful in cases where even the deterministic aspects of the knowledge being modeled is so ill understood that there has been no success in modeling even that part. Putting it another way, the Bayesian framework would be a panacea if the only obstacle to modeling commonsense knowledge wete that it was some probabilistic version of something that could be successfully modeled otherwise. However, we believe that the obstacles are of a different and more severe nature: The basic concepts in this knowledge, as represented typically by natural language words, do not generally have unambiguous meanings. They may number tens or hundreds of thousands, as they do in the experiments reported in [8]. Typically, an observed situation contains much incomplete information - the truth value of most concepts in any one situation is unstated and unknown. Finally, there is no reason to believe that an accurate model of the totality of the possible relationships among these multitudinous concepts exists.

While the predicate calculus and Bayesian reasoning may be useful intellectual aids in designing systems, by themselves they do not offer the guarantee of robustness that is needed: significant aspects of the world are difficult enough to describe accurately, and when conclusions are to be drawn from conjoining a series of these aspects then the errors are likely to grow out of control.

Our proposal is that the only guarantee of robustness that is viable for complex manipulations on uncertain unaxiomatized pieces of knowledge is that offered by learning processes that have access to instances of the world to which the knowledge refers. The knowledge in the system can then be constantly tested and updated against real world examples. The behavior of the system will then be guaranteed in a statistical sense to be correct with high probability on examples drawn from the same probability distribution from which the learning experience was drawn. Thus the semantics we advocate for systems that

manipulate commonsense knowledge is *PAC semantics* [12], which we shall discuss below. We shall require not just the learning aspects but also the *outcomes of the reasoning processes* to be predictably accurate in that sense. An argument for why an *a priori* guarantee of accuracy, as guaranteed by PAC semantics, is needed for all aspects of the system can be illustrated by distinguishing three situations:

In a first kind of situation, which we call (A), we have a candidate intelligent system at hand. To test whether its behavior is effective we can run it on live examples. We will for sure get a reliable statistical assessment of the system's accuracy on the distribution of examples on which it is tested.

In another situation, which we shall label (C), we do not have a candidate system at hand, but are asking whether one can be built at all, and wondering on what principles it might be built so as to be able to pass a live test as described above. The PAC-model of learning is designed exactly for this situation. It promises that a system, based on certain algorithms and trained on enough examples of a fixed but arbitrary function that is within the capabilities of the learning algorithm, will with high probability be able to pass the live test described in situation (A). The PAC model guarantees that the processes are *computationally feasible*, needing only a polynomial amount of computation and data. Equally importantly, the model acknowledges the possibility that errors will be made in the predictions, but these *errors will be controlled* in the sense that they can be made arbitrarily small by increasing, in a polynomially bounded manner, the amount of data and computation that is being invested. The PAC model, which captures and quantifies both the computational and statistical aspects of learning, is designed to capture exactly the desiderata of any system that draws its knowledge from, and needs to perform well in, a world that is too complex to be modeled exactly.

It is conceivable, of course, that systems based on principles, such as Bayesian inference or the predicate calculus, that do not guarantee robustness *a priori* in this way will by chance offer such robustness. This has not happened to date. We would argue that if such robustness is found then that too will be a phenomenon of PAC semantics, and therefore most fruitfully described in those terms. Whatever thought aids may have been used in the design, the only sense in which the result can be declared a success is in the PAC sense that the system is accurate in its ultimate task on natural examples, and requires only efficiently computable processes.

Returning to our enumeration of the different situations, we note that there is also an intermediate situation (B). There we have a candidate system at hand, as in (A), but instead of testing it against live data we are given a set of examples on which the system has performed well, with the promise that the examples were once chosen live from a distribution, but no promise that the system was designed independently of these examples. We can validate the system against the examples, as in (A), but we have reason to be suspicious that the system was tailor made to fit the data. However, ignoring the computational aspects of the PAC model and retaining only the statistical ones, we can obtain confidence in the system if the system is simple enough in terms of the amount of corroborating data, whether this simplicity is measured in terms of the number of bits [1] or the VC-dimension [2] of the system description. This situation (B) is also interesting because it, like situation (A), provides a principled reason for having confidence in a system even if the design methodology of the

system did not guarantee such confidence.

We conclude that what we need ideally is a design methodology that guarantees robustness in the PAC sense, as in situation (C). We may be lucky and derive systems with similar performance in the PAC sense, as verified in situations (A) or (B), without having used a methodology that is guided by a PAC guarantee. However, based on the past history of such attempts, we estimate the likelihood of this succeeding as being small.

We are not suggesting that heuristics, or algorithms whose success is not well understood, be avoided altogether. In robust logic we first learn rules that are accurate in the PAC sense, and then we chain these together in a way that gives predictions that are also accurate in the PAC sense if the learned rules were. It may be valid to use heuristics in each of the two halves if sight is not lost of the overall goal that the final predictions have to be accurate in the PAC sense. For example, the first half is a standard machine learning task. There is ample evidence for the existence of algorithms, such as various decision tree algorithms, that appear to be effective PAC learning algorithms for some useful set of functions and distributions that have yet to be characterized. There is no reason for not using these if these are shown to be effective in practice. What we are saying, however, is that if we do not *plan* for PAC accuracy at every stage, in the manner of robust logic, for example, then we are unlikely to get PAC accuracy in the final predictions.

3 Teaching Materials

The problem of creating systems that realize KI has two parts. The first is the design of the specific learning and reasoning algorithms that are to be used, as discussed for example in [14]. The second is the manner in which the real world knowledge is presented to the system. It may be possible to arrive at reasonable proposals for the former algorithmic questions once and for all. However, the second aspect, which we call the preparation of *teaching materials*, may be an endless task reflective of the endless effort humans put into the analogous process in the education of the young.

While we emphasize that the main characteristic of commonsense knowledge is that no axiomatization is known, we welcome the use of any attempted axiomatizations of parts of the knowledge. For example, when processing natural language texts dictionaries of synonyms and antonyms, as provided, for example, by WordNet [9], are extremely useful, and are used, in fact, in the experiments reported in [8]. Similarly, hand-designed ontologies of knowledge, as developed for example in [5], may have an important role in providing information that is difficult to acquire elsewhere. We shall regard such hand-designed attempted axiomatizations also as teaching materials. When these are used in a KI system they should be regarded as having PAC semantics also, and subject to modification in the light of experience. For example, if a dictionary contains some inconsistencies then this will be discovered in the course of applying this knowledge to examples. Of course, equally welcome as teaching materials to hand-crafted methods, are automatic methods of obtaining reliable knowledge, even when these are of restricted forms (e.g. [3]).

The teaching materials can be expected to have some hand-designed architecture. For example, the knowledge may be layered, so that the most fundamental knowledge is infused first, and subsequent layers that depend on that first layer are infused later. Of course,

the creation of teaching materials for even one layer may be expected to be challenging. Naturally occurring sources, such as books or the web, may omit essential knowledge that humans acquire by other means. We hope that progress in building useful systems will be made, nevertheless, once the problem of constructing teaching materials is raised to the status of a first-class intellectual activity.

A fundamental difficulty may arise in bootstrapping this process. For example, if the lowest layer of concepts on which the multi-layered learning is performed consists of visual primitives, which are at least partially available at birth in biological systems, or of knowledge of three dimensional space at a level not explicitly taught to children, then there remains the problem of providing these primitives to the system. It is conceivable that this can be done by programming. However, there remains the possibility that, just as with higher level concepts, the only practical way of putting these into a machine in a robust enough manner is by learning. Now evolution can also be regarded as a learning process, and recently a theory of evolvability has been formulated in the PAC framework [16]. Hence one can envisage constructing teaching materials for intelligent systems to correspond not only to knowledge learned by individual humans, but also to knowledge acquired by them from their ancestors through evolution. We believe that biology provides an existence proof that cognitive systems based on pure learning and appropriate teaching materials are feasible. It remains, however, a significant research endeavor to find pragmatic ways of constructing useful systems by means of these methods, with or without programmed components.

4 Further Issues

What we have attempted to argue here is that there is no hope of creating intelligent systems if one fails to incorporate mechanisms, in the manner of KI, that guarantee robustness of the decisions made by the system. Over the decades researchers have identified many other difficulties in the pursuit of intelligent systems. The question arises as to whether KI makes some of these difficulties even less tractable, or contributes to alleviating these.

The first general point we make is that, at least from a cognitive perspective, the PAC semantics of KI should be viewed as substantially assumption-free and not as imposing substantive constraints. The definition does presuppose that the function being learned is within the capabilities of the learning algorithm. However, as long as we are learning concepts that are learnable at all, for example by a biological system, then we have an existence proof that such a learning algorithm exists. We note that an actual system will attempt to learn many concepts simultaneously. It will succeed for those for which it has enough data, and that are simple enough when expressed in terms of the previously reliably learned concepts that they lie in the learnable class. The system can recognize which concepts it has learned reliably and which not, and will only use the former for reasoning. In this way a system will have a principled way of discovering which fragments of the knowledge offer useful predictive power, without having to embark on the hopeless task of modeling all of it.

Second, we argue that the statistical notion of correctness against a real world distribution of examples in the PAC sense is the best we can hope for. Of course, in many areas of science, particularly physics, strong predictive models of the world whether deterministic

or probabilistic do hold. This is because these models are based on an axiomatization of a restricted aspect of the world. Clearly, for any aspect of the world that can be similarly axiomatized (i.e. for which an accurate generative model can be designed) whether in terms of differential equations, mathematical logic, or explicit probabilistic models, such models can lead to predictions that work in all cases with quantifiable error and are superior. However, commonsense reasoning addresses areas where such axiomatizations and generative models have met with limited success. In particular systems based on them have not scaled. The considerable success of machine learning as compared with programmed systems, in speech recognition, computer vision and natural language processing, we interpret as deriving from the fact that the robustness that learning offers outweighs the possible benefits of partially correct axiomatizations. For the general commonsense reasoning problem we expect this tradeoff to tilt considerably further towards machine learning.

Finally, we ask whether PAC semantics offers solutions to the difficulties that have been identified for other approaches? This issue has been discussed in [13]. There it is argued that such issues as conflict resolution, context, incomplete information, and nonmonotonic phenomena, which are problematic to various degrees for classical logic, are not inherently problematic in PAC semantics. In fact, interesting new possibilities arise, for example, in the treatment of incomplete information [7].

5 Acknowledgement

I am grateful to Loizos Michael for his helpful comments on this paper.

References

[1] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Occam's razor. *Information Processing Letters*, 24 (1987) 377-380.

[2] A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth. Learnability and the Vapnik -Chervonenkis dimension. *Journal of the ACM*, 36 (1989) 929965.

[3] O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates. Unsupervised named-entity extraction from the web: an experimental study *Artif. Intell.*, Vol. 165, No. 1. (June 2005), pp. 91-134.

[4] Y. Even-Zohar and D. Roth. A classification approach to wod prediction. In *Proc. First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'00)*, (2000) 124-131.

[5] D. B. Lenat. CYC: A large-scale investment in knowledge infrastructure, *Comm. Assoc. Computing Machinery* 38:11 (1995) 33-38.

[6] J. McCarthy. Programs with common sense. In *Proc. Teddington Conference on the mechanization of Thought Processes*, (1958), 75-79.

[7] L. Michael. Learning from partial observations. *IJCAI* (2007), 968-974.

[8] L. Michael and L. G. Valiant. A first experimental demonstration of massive knowledge infusion, *Proc. 11th International Conference on Principles of Knowledge Representation and Reasoning*, Sept. 16-20, 2008, Sydney, Australia, 378-389.

[9] G. A. Miller. WordNet: A lexical database for English. *Comm. Assoc. Computing Machinery*, 38:11(1995) 39-41.
[10] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA (1988).
[11] A. M. Turing. Computing machinery and intelligence. *Mind*, LIX (236):433-460, (1950).
[12] L. G. Valiant. A theory of the learnable. *Comm. Assoc. Computing Machinery* 27:11 (1984) 1134-1142.
[13] L. G. Valiant. A neuroidal architecture for cognitive computation, *J. Assoc. Computing Machinery*, 47:5 (2000) 854-882.
[14] L. G. Valiant. Robust logics, *Artificial Intelligence Journal*, 117 (2000) 231-253.
[15] L. G. Valiant. Knowledge infusion, *Proc. 21st National Conference on Artificial Intelligence, AAAI06*, Jul 16-20, 2006, Boston, MA, AAAI Press, 1546-1551.
[16] L. G. Valiant. Evolvability, *J. Assoc. Computing Machinery*, to appear. (Earlier version: *Proc. 32nd International Symposium on Mathematical Foundations of Computer Science*, Aug. 26-31, Český Krumlov, Czech Republic, *LNCS*, Vol 4708, (2007) Springer-Verlag, 22-43.)