Measuring and Improving the Effectiveness of Representations

Russell Greiner*
Department of Computer Science
University of Toronto
Toronto, Ontario M5S 1A4
greiner@cs.toronto.edu

Charles Elkan†
Department of Computer Science
University of California, San Diego
La Jolla, California 92093-0114
elkan@cs.ucsd.edu

Abstract

This report discusses what it means to claim that a representation is an effective encoding of knowledge. We first present dimensions of merit for evaluating representations, based on the view that usefulness is a behavioral property, and is necessarily relative to a specified task. We then provide methods (based on results from mathematical statistics) for reliably measuring effectiveness empirically, and hence for comparing different representations. We also discuss weak but guaranteed methods of improving inadequate representations. Our results are an application of the ideas of formal learning theory to concrete knowledge representation formalisms.

1 Introduction

A principal aim of research in knowledge representation and reasoning is to design good formalisms for representing knowledge about the world. This paper gives operational criteria for evaluating the goodness of a "representation".[1] Many areas of AI research can use these results. For example, many papers on nonmonotonic logic [Reiter, 1987] implicitly or explicitly make the claim that one formalism leads to *better* representations than the others. Similarly, the aim of an explanation-based learning (EBL) system [Mitchell *et al.*, 1986; DeJong and Mooney, 1986; DeJong, 1988] is to transform one problem-solving representation into a *better* one. However these articles do not specify precisely what it means to say that one representation is better than another. This paper addresses this shortcoming.

Our research has a pragmatic objective: to develop methods for deciding, for example, what type of representation should be installed in Robbie the Robot to enable it to retrieve keys located several miles away. The criteria for Robbie's success are fundamentally behavioral — what is critical is the correctness of Robbie's interactions with the world. This paper therefore adopts an "external" perspective, in which the goodness of a representation depends on its observable outputs.

One alternative perspective is worth mentioning. The "internal" perspective asks questions about a representation, such as whether it is elegant, concise, or simple. These properties may be desirable for certain tasks, such as communicating information to another agent.[2] However, while it is possible to provide operational criteria for evaluating internal properties of a representation, these properties necessarily reflect some subjective choices, and their correlation with the usefulness of a representation is usually difficult to gauge. External criteria, on the other hand, can directly guide the design of useful AI systems.

So, what behavioral features do good representations exhibit? Accuracy and coverage are two: everything else held equal, a representation should state only true facts,

*Supported by an Operating Grant from Canada's Natural Science and Engineering Research Council. Both authors would like to thank William Cohen, Dave Mitchell and the anonymous referees for their useful comments.

†Supported by a grant from the Powell foundation. This research was performed while at the University of Toronto.

[1] "Representation" is an abbreviation for "representational system", which refers to a process that answers questions; see Section 2. *N.b.*, it is not just a data structure, but will include a reasoning process as well.

[2] Much of the work on nonmonotonic formalisms is motivated by internal criteria of goodness. The notorious frame problem, in its simplest guise, is to find a formalism that admits concise representations of actions and their effects; and conciseness is an internal criterion.

and it should state as many true facts as possible. Timeliness is another important property: an accurate answer to a question can be worthless if the answer arrives too late. (The correct information that "stock X will go up tomorrow" is useless unless it arrives before the end of today's trading.) In general we have to settle for representations that are not both universally accurate and fast; *c.f.*, [Nagel and Newman, 1958; Simon, 1981; Levesque and Brachman, 1985]. Any single-valued metric for evaluating representations must embody decisions as to the relative importance of accuracy, coverage, and timeliness. The tradeoffs chosen will depend on the context: for example, fast responses are more important when predicting future events than in most design tasks. This measure must also be relative to the anticipated set of problems. (*E.g.*, it may be acceptable for a representation to return incorrect answers to certain questions, but only if those questions occur very rarely; hence, we need to know how often those questions will appear.)

This report shows how to define measures of the effectiveness of a representation, how to score representations according to these measures, and how to transform representations in ways that increase their scores. Section 2 discusses the dimensions of merit just mentioned and proposes combined utility measures. Sections 3 and 4 show how to use a set of observations to determine with confidence whether a representation is acceptable, and to improve a "deficient" representation, or even to find an approximately optimal representation, with high probability.

2 Framework

Our aim here is not to debate what representations really are; as a working definition, we take it that a representation is some sort of system capable of answering questions about the world. Mathematically, a representation is a function that maps queries to answers. In practice, representations are typically combinations of a data structure encoding aspects of the outside world and an inference procedure; see Section 4 for examples.

Our concern is with the effectiveness of a representation as perceived by an external observer, who wants to use the representation to perform some task. Here, performing a task involves asking the representation questions and receiving its answers. Formally, we assume that there is a set $\mathcal{Q}$ of all relevant questions that the representation may be asked.

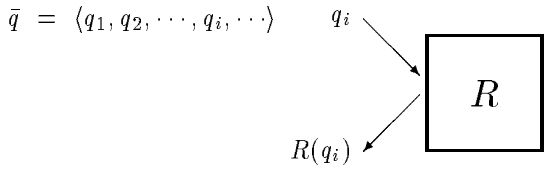

Figure 1: General view of a representation.

In Figure 1, R is the name of a representation that takes as input a sequence of queries $\bar{q} = \langle q_1, q_2, \ldots, q_i, \ldots \rangle$ and after each query q_i, produces a candidate answer $R(q_i)$. Each q_i is drawn randomly from $\mathcal{Q}$ according to an arbitrary, stationary probability distribution.

For simplicity we shall assume that the response to any query is one of {Yes, No, IDK}, where IDK is read as "I don't know." In order to talk about the accuracy and coverage of representations, we need a specification of when responses are correct. Formally, we posit the existence of an oracle, $\mathcal{O}(\cdot)$, that maps each query onto its unique correct answer Yes or No.[3] Finally, to discuss the speed of a representation, we need a time function $\tau(R, q_i)$ that returns the time required by R to respond to the query q_i.[4]

One example of a representation is a set of axioms in any standard logic, together with an appropriate inference procedure. If the collection of axioms is insufficient, or if the inference procedure is incomplete, then the representation's response to some queries may be IDK. If some of the axioms in the set are false, or if the inference procedure is unsound, then some Yes or No answers may be incorrect.[5]

Dimensions of merit: Criteria for the effectiveness of a representation depend on the expected use of the representation. Some tasks require complete accuracy, while others can allow some inaccuracy in exchange for increased efficiency, and so on. Nevertheless, it is possible to identify some orthogonal dimensions of merit.

[3] The existence of $\mathcal{O}(\cdot)$ presupposes that queries have unique answers. This condition could be relaxed. In addition, the answers to questions could also provide extra information, such as a witness for an existential query, or information about the derivation path used to solve a problem. See [Greiner and Elkan, 1991].

[4] We could use the τ function to deal with resource requirements in general, including space usage, power consumption, etc., as well as time.

[5] If the logic used by the representation has a model-theoretic semantics, the notional oracle that defines the correct answers of queries can be identified with an intended model.

With respect to the query q, a representation R is

accurate	if it responds with either the correct answer $\mathcal{O}(q)$ or IDK when asked q: *i.e.*, if $R(q) \in \{\mathcal{O}(q), \text{IDK}\}$
categorical	if it responds either Yes or No but not IDK: *i.e.*, if $R(q) \in \{\text{Yes}, \text{No}\}$
t-**efficient**	if, when asked q, it responds in at most t time units: *i.e.*, if $\tau(R, q) \leq t$.

These three properties are orthogonal, in the sense that a representation can possess any combination of them. As a simple example, the "ignorant" representation that immediately answers every question with IDK, is universally accurate and ϵ-efficient for some small ϵ, but it is sadly uncategorical.

The three dimensions of merit listed above are not exhaustive: *i.e.*, they do not span the space of external criteria of merit for representations. In particular, [Segre, 1988] identifies a number of further dimensions of merit that are important if the world being modeled by the representation is uncertain. Our dimensions are still quite general; we show below that linear combinations cover many standard situations.

Utility measures: A utility measure evaluates a representation with respect to the set $\mathcal{Q}$ of all questions it will be asked. Taking into account the probability distribution over $\mathcal{Q}$, the measure $\mathcal{M}(R)$ of a representation R is defined as the expected value of $g_{\mathcal{M}}$ over $\mathcal{Q}$, where $g_{\mathcal{M}}$ is a function that returns the score of R on a single query; the arguments of $g_{\mathcal{M}}$ are the query, its correct answer, the representation's answer to the query, and the time required. If $\bar{q}$ is an infinite sequence of queries, such that each element of $\bar{q}$ is drawn independently from $\mathcal{Q}$ according to the fixed probability distribution, then $\mathcal{M}(R)$ can also be defined by the following equation:

$$\mathcal{M}(R) \;=\; \lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} g_{\mathcal{M}}(\, q_i,\, R(q_i),\, \mathcal{O}(q_i),\, \tau(R, q_i)).$$

We shall only consider utility measures whose scoring functions are bounded above and below; *i.e.*, in the class U_λ for some $\lambda \in \mathcal{R}^+$, where

$$U_\lambda = \left\{ g_{\mathcal{M}} \mid \begin{array}{l} \exists a \in \mathcal{R}^+ \; \forall R \; \forall q \in \mathcal{Q}. \\ a \;\leq\; g_{\mathcal{M}}(\, q_i,\, R(q_i),\, \mathcal{O}(q_i),\, \tau(R, q_i)) \;\leq a + \lambda \end{array} \right\}$$

This property is technically convenient for proving the results of the following sections, but similar results hold for the class U of all utility measures.

Since a perfectly accurate, categorical, and efficient representation is usually unobtainable, and the relative importance of the three dimensions depends on the task for which the representation is to be used, we now define a class of utility measures that incorporate task-specific tradeoffs. This class consists of weighted linear combinations of an accuracy value, a categoricity value, and a timeliness value; as shown below:

$$g(q, R(q), \mathcal{O}(q), \tau(R, q)) \quad = \quad \left\{ \begin{array}{ll} \alpha_+ & \text{if } R(q_i) = \mathcal{O}(q_i) \\ -\alpha_0 & \text{if } R(q_i) = \text{IDK} \\ -\alpha_- & \text{otherwise} \end{array} \right\} \; - \; \alpha_t \cdot \tau(R, q_i).$$

This equation characterizes the entire class of bounded, linearly separable utility measures, where the time penalty is linear.[6] There are several important special cases: Setting $\alpha_t = 0$ indicates that efficiency is unimportant; setting $\alpha_0 = \alpha_-$ discriminates only between correct and incorrect answers, etc.

3 Appraising and Comparing Representations

Appraising utility: Suppose someone hands you a black-box representation and claims that the $\mathcal{M}$ value of this representation is above some threshold. How confident should you be with that assessment? There are clearly several approaches. First, you might trust the supplier of the representation — some people really do think they are purchasing the Brooklyn Bridge. Second, you could analyze the internals of the representation. For example, imagine that the representation's answers are the result of running a sound proof procedure on a collection of true facts. You can then be completely confident that all its answers will be accurate. If you know, furthermore, that all facts have a certain syntactic form (*e.g.*, all are DATALOG definite clauses), you can know that all answers are obtainable in polynomial time. The remainder of this section presents a third approach, in which a sequence of examples of the representation's behavior is used to approximate its real utility.

Our objective is to determine the worth of a representation *for some specific application*; that is, we want to know how it will perform over the particular distribu-

[6] We assume that $\tau(R, q)$ "tops off" at some upper limit. Notice this assumption holds for all of the utility measures discussed in Section 4.

tion of queries it will eventually see. The only empirical method that is guaranteed to provide this information involves observing the representation perform throughout its entire lifetime; a most impractical approach. Fortunately, however, we can obtain good estimates, based on a small set of samples, using "Chernoff bounds" [Chernoff, 1952]:

> Let $\{X_i\}$ be a set of independent, identically distributed random variables with mean μ, whose values are all in the range $[a, a+\lambda]$; and let $S_n = \frac{1}{n}\sum_{j=1}^{n} X_j$ be the sample mean of n variables. We expect this average to tend to the population mean, μ, as $n \to \infty$. Chernoff bounds tell us the probable rate of convergence: the probability that "S_n is more than $\mu+\beta$" goes to 0 exponentially fast as n increases; and, for a fixed n, exponentially as β increases. Formally, [Bollobás, 1985, p. 12]
>
> $$Pr[\ S_n > \mu + \beta\ ] \ \leq\ e^{-2n\left(\frac{\beta}{\lambda}\right)^2} \qquad (1)$$

Recall now that a representation's utility is defined as the mean of its scores over its distribution of possible queries. Now suppose that the value of the scoring function is always in the range $[-3, -3+5]$, say, and that the average observed score of the representation R is $S_{1000} = 1.7$ after $n = 1000$ queries. Equation 1 says that the probability that the representation's real utility is greater than, say, $\beta = 0.25$ over the real utility is less than $e^{-2n(\beta/\lambda)^2} = e^{-2\times 1000\times(\frac{0.25}{5})^2} \approx 0.0067$.

We can therefore use the performance of a representation on a sampling of queries to approximate the representation's true utility. The following lemma supplies the sample complexity required to estimate a representation's utility to high accuracy (within ϵ) with high probability (at least $1-\delta$).[7]

Lemma 1 *Let ϵ, $\delta > 0$ be given constants, R be a representation, $\mathcal{M}(\cdot) \in U_\lambda$ be a utility measure, and $\hat{\mathcal{M}}^{N_1}(R)$ be the approximation to $\mathcal{M}(\cdot)$ obtained as the average score on $N_1 \stackrel{def}{=} \lceil \frac{1}{2}\left(\frac{\lambda}{\epsilon}\right)^2 \ln\frac{2}{\delta}\rceil$ sample queries. Then*

$$Pr\left[\ |\mathcal{M}(R) - \hat{\mathcal{M}}^{N_1}(R)| \ \leq\ \epsilon\ \right] \quad \geq \quad 1-\delta.$$

This means we can approximate the representation's true utility, with provably high probability to arbitrarily high accuracy, by watching its behavior over enough samples. Notice the estimate $\hat{\mathcal{M}}^{N_1}(R)$, like the real $\mathcal{M}(R)$, depends on the distribution of queries that R will encounter; it therefore is more likely to provide a good indication of R's true utility than we would get by testing R on worst case queries, a set of concocted queries, a sample drawn randomly from a uniform distribution [Goldberg, 1979], or any particular collection of "benchmark challenge problems" [Keller, 1987].

Comparing representations: If we had an analytic technique for evaluating the utility of representations with respect to a distribution of queries, and knew this distribution of queries, we could directly determine which of two representations was better. In general we have neither analytic technique nor distribution, but we can fall back on an empirical technique — of "running" both contenders; *i.e.*, finding the "paired-t confidence" [Law and Kelton, 1982].

Lemma 2 *Let $\delta > 0$ be a given constant, and $\mathcal{M}(\cdot) \in U_\lambda$ be a utility measure. Given any two representations, R_1 and R_2, let $\hat{\mathcal{M}}^N(R_1)$ (respectively $\hat{\mathcal{M}}^N(R_2)$) be the approximation to $\mathcal{M}(R_1)$ (respectively $\mathcal{M}(R_2)$) obtained as the average score on $N \in \mathcal{Z}^+$ samples. If*

$$\hat{\mathcal{M}}^N(R_2) - \hat{\mathcal{M}}^N(R_1) \quad > \quad \lambda\sqrt{\frac{2}{N}\ln\frac{1}{\delta}}$$

then $\mathcal{M}(R_2) > \mathcal{M}(R_1)$ holds with confidence $\geq 1-\delta$. (Hence, we may believe that R_2 is $\mathcal{M}$-better than R_1).

4 Improving a Representation

This section discusses various ways to improve a deficient representation. The lemmas of the previous section directly suggest a "generate and test" procedure for improving representations. Other approaches search the space of alternative representations more intelligently. They use a set of "training examples" (perhaps $\langle q_i, \mathcal{O}(q_i)\rangle$ query/answer pairs) to construct a new representation that is $\mathcal{M}$-better than the original, with provably high probability.

Due to space limitations, this short paper can only summarize a few individual algorithms associated with various type of utility measures. The extended paper, [Greiner and Elkan, 1991], presents these algorithms (and others) in detail, and proves their correctness.

Improving efficiency (preserving accuracy): Suppose that a representation is producing the correct answers, but too slowly. As an example, consider the knowledge base shown in Figure 2, that states that we will buy a car if it is fast or cheap. Furthermore, our

[7] [Greiner and Elkan, 1991] provides the proofs for all of the lemmas and theorems presented in this short paper.

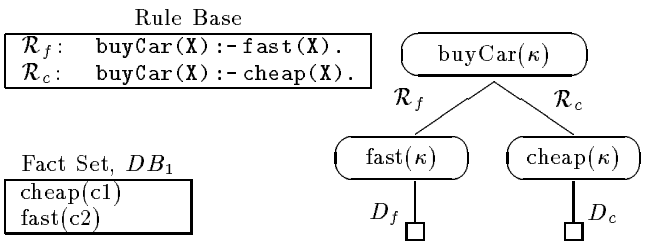

Figure 2: Rule base and associated inference graph

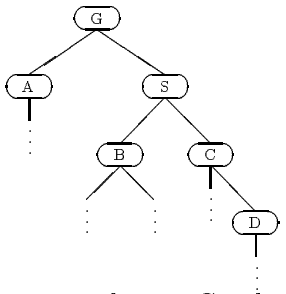

Figure 3: Inference Graph, G_M

representation R_1 searches this "inference graph" in a left-to-right order — *i.e.*, given a query `buyCar(c`$_i$`)` (for some constant `c`$_i$), it would first use the $\mathcal{R}_f$ rule to reduce `buyCar(c`$_i$`)` to the subgoal `fast(c`$_i$`)`, and then determine whether this proposition is in the fact set, DB_1. If so (for example in the case where the top query was `buyCar(c2)`), R_1 would return `Yes` and terminate. If not, it would then follow the $\mathcal{R}_c$ rule to the `cheap(c`$_i$`)` subgoal, and ask if it is in the database. R_1 would then return `Yes` or `IDK`, depending on the success of that retrieval. Here, we say that R_1's *strategy* is $\langle \mathcal{R}_f, D_f, \mathcal{R}_c, D_c \rangle$.

Now imagine we find that R_1 is relatively slow — perhaps because the queries all dealt with `c1`, meaning R_1 "wasted" its time on the $\langle \mathcal{R}_f, D_f \rangle$ path, before finding the appropriate $\langle \mathcal{R}_c, D_c \rangle$ path. One obvious proposal would be to change R_1 to a new representation, R_2, that uses the different strategy, $\langle \mathcal{R}_c, D_c, \mathcal{R}_f, D_f \rangle$. Notice this R_2 would be strictly better than R_1, if all of the queries were `buyCar(c1)`. In general, we can compute which the relative utilities of R_1 and R_2 if we know the costs of the various paths (*e.g.*, the cost of following the $\langle \mathcal{R}_c, D_c \rangle$ path), and the frequencies of the various possible queries (*e.g.*, that 80% of the queries would be `buyCar(c1)`, 5% would be `buyCar(c2)`, and 15% would be `buyCar(c)` for some `c` such that `buyCar(c)` is not provable; see [Greiner, 1991]).

In general, we do not know this distribution information. We can, however, use a set of samples to estimate it, and use this information to determine whether R_2 would be better than R_1, with high confidence. In fact, we can obtain reasonable estimates of this distribution by running R_1 *alone*; see [Greiner and Cohen, 1991a]. That paper also shows how to extend this learning system to handle more elaborate classes of inference graphs (with many different paths from query to retrieval, involving conjunctions, etc.), and other types of modifications (besides simply rearranging the strategy). It also presents an efficient algorithm, PALO, that can "hill-climb" in the space of representations, reaching a representation that is, with high probability, close to a local optimum. The PAO system [Greiner and Orponen, 1991] is another variant of this idea: it uses a set of training examples to identify a strategy that is, with high probability, arbitrarily close to the globally optimal strategy.

Increasing efficiency by decreasing accuracy: The previous subsection deals with speedup learning (performing only "symbol-level" modifications [Dietterich, 1986]), whose objective is new performance systems that are as accurate and as categorical as the original system, but (we hope) faster. This subsection removes this constraint, and considers techniques that are allowed to produce systems that are less accurate or less categorical. The utility-based approach introduced in this paper allows us to quantify the tradeoffs among the various performance attributes of a representation. In particular, the relative values of α_+ and α_- versus α_t specify how much accuracy or categoricity we are willing to lose, in exchange for how much improvement in speed.

As a simple example of an application of this idea, consider the G_M inference graph shown in Figure 3. As above, each node represents a (sub)goal, and each arc, a rule that reduces one goal to a subgoal. Some arcs are "probabilistic", meaning that whether a representation can traverse such an arc can depend on the contents of the fact set, or the particular query posed, etc. A representation returns `Yes` if it reachs any leaf node. (More details appear in [Greiner and Cohen, 1991a].)

The R_1 representation answers G($\cdot$) queries by exploring the entire G_M graph. Imagine the subgraph under the B node is extensive, meaning it is expensive to search, and furthermore, that R_1 seldom find answers in that subgraph, over the distribution of G($\cdot$) queries.

Now consider the R_2 representation, which traverses

some of the arcs in G_M, in the same order as R_1; but R_2 completely skips B and its children. Of course, this R_2 will not reach any node under B, and so can produce an incorrect answer. It will, however, require less time to produce that answer. Is it better than R_1? This depends on the relative values of α_t versus $\alpha_+ + \alpha_0$, and on the actual cost of searching under B, versus the frequency with which solutions will appear there.

We can consider this "delete subgraph" process as an operation. As above, we can provide conditions under which it (is likely to) produce a superior new representation, and build a PALO-like learning system that hill-climbs in this space to a representation that is, with arbitrarily high probability, close to local optimum.

Two final notes: [1] We can use this same approach to deal with SAT problems: here, the inference graph is a complete binary tree, where each node corresponds to a literal, and one descending arc indicates that this literal should be positive, and other, that is should be negative. [2] Several previous systems have implicitly dealt with this theme, of gaining efficiency at the expense of accuracy or categoricity; see [Cohen, 1990], [Subramanian and Genesereth, 1987], [Levesque, 1986; Selman and Kautz, 1988; Etherington *et al.*, 1989; Borgida and Etherington, 1989], [Selman and Kautz, 1991]. None, however, have explicitly quantified the tradeoffs.

Improving accuracy (ignoring efficiency): One standard issue with default logic is the "multiple extension problem" [Reiter, 1987]: For example, knowing the facts $\mathcal{F}_1 = \{\mathrm{B(T)}, \mathrm{P(T)}\}$ and the defaults $\mathcal{D}_1 = \left\{ d_1 : \frac{\mathrm{B(x)}:\ \mathrm{F(x)}}{\mathrm{F(x)}},\ d_2 : \frac{\mathrm{P(x)}:\ \neg\mathrm{F(x)}}{\neg\mathrm{F(x)}} \right\}$, one could conclude either F(T), based on the first fact and first default, d_1; or $\neg$F(T), based on the second fact and second default, d_2.[8]

One way around this problem involves prioritizing the defaults [Przymusiński, 1987; Brewka, 1989]. Here, for example, we could specify $d_2 \prec d_1$, meaning we should use the d_2 default if it applies, and only if it does not should we consider d_1. Hence, R_1, the default-based representation that uses this ordering, would conclude $\neg$F(T) as desired, and not F(T).

Unfortunately, there is not always an appropriate order of the defaults; consider, for example, the famous "Nixon diamond", based on the facts $\mathcal{F}_2 = \{\mathrm{Q(N)}, \mathrm{R(N)}\}$ and the defaults $\mathcal{D}_2 = \left\{ d_3 : \frac{\mathrm{Q(x)}:\ \mathrm{P(x)}}{\mathrm{P(x)}},\ d_4 : \frac{\mathrm{R(x)}:\ \neg\mathrm{P(x)}}{\neg\mathrm{P(x)}} \right\}$. Here, one could either use d_3 to conclude P(N), or use d_4 to conclude $\neg$P(N).[9]

Which ordering is appropriate: $d_3 \prec d_4$ or $d_4 \prec d_3$? Our paper takes a pragmatic approach to answering this question: the proper order is the one that produces the correct answer most often. This, of course, depends on the distribution of queries. For example, imagine we knew that the only individual that mattered (*i.e.*, the only n such that Q(n) and R(n) held, and that appeared in any "P(n)?" query) was $\mathtt{n}_1$, and that this $\mathtt{n}_1$ was a pacifist — *i.e.*, the oracle (that emulates the "real world") claims P(n_1). The "pragmatically correct" ordering, here, is $d_3 \prec d_4$, as this leads to the correct answer. Similarly, if we knew that there were many relevant individuals, but very few were pacifists, then we would know that $d_4 \prec d_3$ is appropriate.

In general, however, we do not know the distribution. [Greiner and Cohen, 1991b] presents a PAO-like algorithm that uses a set of samples to approximate the distribution, to obtain an estimate that is sufficiently close to guarantee that the total ordering based on it will be close to the global optimal, with high probability. It also shows this task to be NP-hard, even in the simple case when only a single default is used in any derivation. That paper then discusses a PALO-like hill-climbing system that finds a representation that is usually arbitrarily close to a local optimum.

We close with a few final comments. [1] Notice we are only concerned with the *accuracy* of the resulting representation, but not with its efficiency. Hence, this learning process corresponds to a utility measure with $\alpha_t = 0$. [2] The "accuracy" of the resulting representation is improved with respect to the $\mathcal{O}$ oracle. While we think of this oracle as an encoding of the real world, this is not necessary; this overall learning system can be viewed as a way of obtaining a increasingly more correct "simulation" of an arbitrary function. [3] There are other approaches to this challenge. One obvious one is to remove the defaults (here usually called "hypotheses", etc.) that are inconsistent with the data. Essentially all inductive inference systems fit into the framework; see

[8] B = "Bird"; P = "Penguin"; T = "Tweety"; F = "Fly".

[9] N = "Nixon"; Q = "Quaker"; R = "Republican"; P = "Pacifist".

[Russell and Grosof, 1987], [Haussler, 1988].

5 Conclusion

The principal claim of many papers on knowledge representation is that one proposed representation is appropriate, or that one is better than another: *e.g.*, that this axiomatization of liquids is good, or that this EBL process does produce improved problem-solvers, and so on. The work reported here provides a way for future papers to add teeth to such claims, by defining a general approach to evaluating and comparing representations. The approach is based on the position that whether a representation is appropriate, or better than another, is a *behavioral* property, that must be defined relative to a specified task. In this paper, tasks are modelled as distributions of anticipated queries, and the desired behavior of a representation is specified using a scoring function. In the spirit of traditional scientific methodology, the notion of representation usefulness is given an operational definition that has no dependence on any subjective observer.

After settling on a precise definition of usefulness, this paper provides a method for reliably (but approximately) measuring usefulness, and presents several transformations capable of improving an inadequate representation. The space of techniques for improving representations is open-ended. We hope other researchers will formulate their mechanisms in this framework, and so contribute to the growing class of useful transformations.

References

[Bollobás, 1985] B. Bollobás. *Random Graphs*. Acad. Press, 1985.

[Borgida and Etherington, 1989] A. Borgida and D. Etherington. Hierarchical knowledge bases and efficient disjunctive reasoning. In *KR-89*, 1989.

[Brewka, 1989] G. Brewka. Preferred subtheories: An extended logical framework for default reasoning. In *IJCAI-89*, 1989.

[Chernoff, 1952] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sums of observations. *Annals of Mathematical Statistics*, 23, 1952.

[Cohen, 1990] W. Cohen. Using distribution-free learning theory to analyze chunking. In *CSCSI-90*, 1990.

[DeJong and Mooney, 1986] G. DeJong and R. Mooney. Explanation-based learning: An alternative view. *Machine Learning*, 1(2), 1986.

[DeJong, 1988] G. DeJong. AAAI workshop on Explanation-Based Learning. Sponsored by AAAI, 1988.

[Dietterich, 1986] T. Dietterich. Learning at the knowledge level. *Machine Learning*, 1(3), 1986.

[Etherington *et al.*, 1989] D. Etherington, A. Borgida, R. Brachman, and H. Kautz. Vivid knowledge and tractable reasoning: Preliminary report. In *IJCAI-89*, 1989.

[Goldberg, 1979] A. Goldberg. An average case complexity analysis of the satisfiability problem. In *Proceedings of the 4th Workshop on Automated Deduction*, 1979.

[Greiner and Cohen, 1991a] R. Greiner and W. Cohen. EBL systems that (almost) always improve performance. Technical report, Univ. of Toronto, 1991.

[Greiner and Cohen, 1991b] R. Greiner and W. Cohen. Producing more accurate theories. Technical report, Univ. of Toronto, 1991.

[Greiner and Elkan, 1991] R. Greiner and C. Elkan. Effective representations. Technical report, Univ. of Toronto, 1991.

[Greiner and Orponen, 1991] R. Greiner and P. Orponen. Probably approximately optimal derivation strategies. In *KR-91*. Morgan Kaufmann, 1991.

[Greiner, 1991] R. Greiner. Finding the optimal derivation strategy in a redundant knowledge base. *Artificial Intelligence*, 1991.

[Haussler, 1988] D. Haussler. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial Intelligence*, 1988.

[Keller, 1987] R. Keller. Defining operationality for explanation-based learning. In *AAAI-87*, 1987.

[Law and Kelton, 1982] A. Law and W. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill Book Co., 1982.

[Levesque and Brachman, 1985] H. Levesque and R. Brachman. A fundamental tradeoff in knowledge representation and reasoning. In *Readings in Knowledge Representation*. Morgan Kaufmann Publishers, Inc., 1985.

[Levesque, 1986] H. Levesque. Making believers out of computers. *Artificial Intelligence*, 30(1), 1986.

[Mitchell *et al.*, 1986] T. Mitchell, R. Keller, and S. Kedar-Cabelli. Example-based generalization: A unifying view. *Machine Learning*, 1(1), 1986.

[Nagel and Newman, 1958] E. Nagel and J. Newman. *Gödel's Proof*. New York University Press, 1958.

[Przymusiński, 1987] T. Przymusiński. On the declarative semantics of stratified deductive databases and logic programs. In *Foundations of Deductive Databases and Logic Programming*. Morgan Kaufmann Pub., Inc., 1987.

[Reiter, 1987] R. Reiter. Nonmonotonic reasoning. In *Annual Review of Computing Sciences*, volume 2. Annual Reviews Incorporated, 1987.

[Russell and Grosof, 1987] S. Russell and B. Grosof. A declarative approach to bias in concept learning. In *AAAI-87*, 1987.

[Segre, 1988] A. Segre. Operationality and real-world plans. In *AAAI Workshop on Explanation-Based Learning*, 1988.

[Selman and Kautz, 1988] B. Selman and H. Kautz. The complexity of model-preference default theories. In *CSCSI-88*, 1988.

[Selman and Kautz, 1991] B. Selman and H. Kautz. Knowledge compilation using horn approximations. In *AAAI-91*, 1991.

[Simon, 1981] H. Simon. *The Sciences of the Artificial*. M.I.T. Press, 1981.

[Subramanian and Genesereth, 1987] D. Subramanian and M. Genesereth. The relevance of irrelevance. In *IJCAI-87*, 1987.